NBA Player Statistics Analysis

By Ryan Kempner

Introduction

The evolution of the NBA (National Basketball Association) has been driven by dramatic changes in player roles, team strategies, rule modifications, and the league's overall style of play. From the dominant big men of the early decades like Shaq to the positionless, three-point-focused offenses of today headlined by Steph Curry and the Golden State Warriors, the league has seen continual transformation. With the ability to access every NBA season and historical records, I can now examine these shifts through statistical analysis. This capstone project explores long-term trends in NBA player performance using a comprehensive dataset containing player per-game statistics from 1947 to 2024.

Basketball fans often rely on subjective opinions or highlight reels to compare players across generations. I aim to use statistical tools to uncover trends in player performance over time. This includes understanding how player archetypes have evolved, what metrics define all-time great seasons, and how MVP-caliber contributions can be predicted with data.

When talking about player archetypes, it is important to understand that within the NBA, and basketball in general, different positions on the court have different jobs, and characteristics that make players better suited for specific positions. The three main positions being focused in this paper are guards, smaller guys who quarterback the offense and touch the ball the most. Centers who are taller guys that are down low defending the basket, and forwards who are a balance of both. However, there are many different archetypes for each position. For example, for guards you can have a pass first guard or a guard who is a three point shooter, or a guard who is a defensive specialist. For forwards, you can have a two way forward who plays offense and defense or a wing forward who is more of a 3 point shooter. For centers, you have the old school dominant center who played down low and blocked shots and only dunked the basketball. Nowadays you have centers that do it all by occasionally shooting the three ball like a guard or forward.

From Kaggle I was able to get a data set that had every single player's counting stats for every season from 1947-2024. In this dataset it listed every player's PPG (points per game) which is how many points a player got a game on average per game. Points are scored by putting the ball in the basket. APG (Assists per game). An assist is when a player passes the ball to another player, who shoots the ball and scores. This stat measures every player's assists on average per game. RPG (rebounds per game). A rebound is when a player shoots the ball and misses and somebody grabs the ball after the miss. This stats measures how many rebounds on average per game someone got. SPG (steals per game) when someone takes the ball from an opposing player. This stats measures how many blocks on average per game. Also BPG (blocks per game). Blocks is when a player contests an opposing team's shot and is able to deflect the player's shot away from the basket. This stat measures on average how many blocks per game a player gets. This dataset had over 55,000 players. I was able to clean this dataset to 13,000 by taking out players who played less than 3/5th of games a season and players who did not have high enough total counting stats to be relevant. I also took out years like the 2020 covid year during which the NBA did not play a full season leading to skewed stats.

This project addresses several key questions about NBA player performance and evolution. I begin by analyzing how core statistics—points, rebounds, assists, steals, and blocks—have changed over the decades, both overall and by position. I then explore how scoring inequality has varied over time using the Gini coefficient. This provides insight into whether today's scoring is more evenly distributed or skewed towards high caliber players. I also identified historically dominant seasons using combined Z-scores across key stats, highlighting the most impactful individual performances. Next, I examine which statistical traits predict longer NBA careers through a multiple regression model. I also tested whether player stats significantly differ across positions using ANOVA. Finally, I develop a logistic regression model to estimate the probability of a player becoming an MVP finalist based on their per-game performance. Together, these analyses provide a comprehensive look at how player roles, statistical impact, and greatness have evolved in the NBA.

Hypothesis

I hypothesize that NBA player performance has evolved toward greater versatility over time, with modern players contributing across multiple statistical categories more consistently than in earlier eras.

Additionally, I expect that: - Traditional position-based differences in player stats have become less pronounced over time, - Outlier seasons will cluster around certain key eras (such as the 1960s and 2010s), - Players with high all-around performance will have longer careers, - And MVP-caliber seasons can be predicted using a combination of basic per-game statistics.

How Have Player Stats Evolved Over Time?

How have key performance metrics such as points, rebounds, assists, steals, and blocks changed over the decades? How have roles shifted by position?

Interpretation:

The line plot **(Figure 1)** highlights how average player performance in key statistical categories has shifted across NBA decades. I see that **scoring (PPG)** peaked during the 1960s and early 1970s, followed by a dip in the late 1990s and early 2000s, and then a resurgence in recent years — reflecting the modern offensive focus and a more spaced out fast paced style of play.

Rebounds (RPG) and **blocks (BPG)** saw a noticeable decline after the 1980s, due to changes

in defensive rules, like goaltending and not being able to stand in the paint. Could also be from fewer

post-oriented players, and a faster game that relies more on 3pt shooting. Meanwhile, assists (APG) have remained relatively stable but have slightly increased recently, driven by more ball movement and team-oriented offenses.

Interestingly, **steals (SPG)**, a stat not officially recorded until the 1970s, have remained consistent but lower in magnitude compared to other stats. This overall trend suggests a transition from interior-dominated play to a more fluid, perimeter-based game, driven by analytics and rule evolution

How Have Player Stats Evolved Over Time?

Are high scorers also strong passers or defenders? What combinations of stats tend to occur together?

Interpretation:

The faceted line plots **(Figure 2)** show how player stats — points (PPG), rebounds (RPG), and assists (APG) — have evolved across decades by simplified position: Guard, Forward, and Center.

- **Guards** have consistently led in assists (APG), with a clear rise over time, reflecting their increasing playmaking responsibilities in modern offenses.
- **Centers** historically dominated in rebounds (RPG) and were high scorers in earlier decades, but their PPG and RPG have declined with the league's shift toward perimeter play.
- **Forwards** sit between guards and centers in both scoring and rebounding, showing a more balanced stat line. Their assist numbers have also slightly increased, suggesting a broader role in offense.

These trends reflect the NBA's move toward more fluid, versatile roles. Especially as guards take on more scoring, and forwards contribute more playmaking, while traditional center dominance fades.

How Is Scoring Distributed Across the League?

Has scoring become more balanced or increasingly concentrated among high usage players?

Interpretation:

The Gini coefficient measures scoring inequality across the league. A value closer to 0 indicates evenly distributed scoring among players, while a value closer to 1 indicates that a few players are responsible for most of the scoring.

From the **Figure 3**, I observe from the graph that:

- Scoring inequality was relatively high in the 1950s and early 1960s This was due to players like Wilt Chamberlain who averaged around 50 points per game. I also had players such as Oscar

Robertson who averaged around 30 points per game. The games were also faster paced meaning teams had more possessions per game as teams averaged around 115 points per game.

- During the 1970s, the Gini coefficient decreased, indicating a more balanced scoring distribution across the league. This is because the NBA added rules to limit players like Wilt Chamberlain
- Starting in the 1980s and into the early 2000s inequality rose again, possibly due to the emergence of high-usage players like Kobe Bryant and isolation-heavy offensive strategies.
- In more recent years, particularly since the 2010s, the Gini coefficient has generally declined, suggesting that scoring has become more balanced and team-oriented — a result of fast paced spread out games, three-point shooting, and improved ball movement.

Overall, the data reflects how the distribution of scoring talent has evolved alongside changes in NBA style and strategy

What Do Outlier Seasons Reveal?

Which player seasons stand out statistically, and what do they tell us about individual greatness?

Interpretation:

The scatter plot **(Figure 5)** identifies extreme outlier seasons based on a combined Z-score across points, rebounds, and assists per game. Higher Z-scores indicate statistical seasons far above the league average in all three areas. Notable observations:

- **Wilt Chamberlain** dominates the left half of the chart, with multiple outlier seasons from the early 1960s — including several with Z-scores above 4.0. His statistical dominance across scoring and rebounding is unmatched in league history.
- **Oscar Robertson** appears frequently in the 1960s as well, especially due to his triple-double level contributions in all three stat categories.
- In the modern era, **Russell Westbrook** stands out with multiple seasons between 2016–2022, reflecting his historic triple-double campaigns.
- Other modern stars such as **Luka Dončić, James Harden, Nikola Jokić**, and **LeBron James** also show up as recent statistical outliers.

The presence of players from both older and current eras demonstrates how statistical greatness can take different forms across time, but multi-category dominance like averaging a triple double remains the main reason for elite outlier seasons.

What Factors Predict a Longer NBA Career?

Are players who score more likely to have longer careers, or does versatility in rebounding and passing matter more? In this section, I build a multiple linear regression model to predict career length based on a player's average points, rebounds, and assists per game.

By identifying which stats are most strongly associated with longevity, I can better understand what traits contribute to extended NBA success beyond just scoring alone.

Interpretation:

The scatter plot **(Figure 7)** compares each player's actual career length (y-axis) to the length predicted by the regression model (x-axis), which is based on their average points, rebounds, and assists per game.

- The dashed red line represents a perfect prediction — points falling directly on this line indicate players whose careers were predicted exactly by the model.
- Most points cluster fairly close to the line, especially for careers between 2 and 10 seasons, suggesting the model performs reasonably well in this range.
- A few players fall far above the line, meaning they had **longer careers than expected** given their stats — these could be role players.
- Some players fall below the line, indicating **shorter-than-expected careers**, potentially due to injuries or off-court factors.

The model effectively captures broad trends in how statistical performance relates to career longevity.

Do Player Stats Significantly Differ by Position?

How do average player statistics vary between guards, forwards, and centers? In this section, I use ANOVA and Tukey post-hoc tests to determine whether there are statistically significant differences in performance based on primary position, across key stats like points, rebounds, assists, steals, and blocks.

Interpretation:

The ANOVA results seen in **Figure 8** indicate that player stats differ significantly by position for all five categories: points (PPG), assists (APG), rebounds (RPG), steals (SPG), and blocks (BPG), with all p-values < 0.001.

The post-hoc Tukey tests reveal consistent trends:

- **Guards** average more assists and steals than forwards and centers, reflecting their ball-handling and perimeter defensive roles.
- **Centers** dominate in rebounds and blocks, which aligns with their traditional role in the basket.

- **Points per game (PPG)** is highest on average for guards, followed by forwards, with centers scoring the least — due to modern offenses emphasizing guard and wing scoring.

These results confirm long-standing positional archetypes in basketball and provide statistical evidence that player responsibilities — and resulting stat lines — are meaningfully shaped by their primary role on the court.

Can I Predict MVP Finalists Using Stats Alone?

Using past data and logistic regression, can I predict the most likely MVP candidates for 2024?

Interpretation:

This logistic regression model **(Figure 10)** estimates the probability of a player being an MVP finalist based on their per-game statistics in points, rebounds, assists, steals, blocks, and minutes played.

Key takeaways from the model summary:

- **All six predictors are statistically significant** ($p < 0.001$), indicating each stat contributes to predicting MVP finalist likelihood. • The largest coefficients are for **assists (APG)** and **minutes per game (MPG)**, suggesting that playmaking and heavy usage are strong indicators of MVP impact.

- **Points (PPG)** also has a large, positive coefficient, as expected — scoring still plays a crucial role in MVP candidacy.

The model was applied to the 2024 season data to generate finalist probabilities. The top 5 predicted MVP candidates include:

1) **Luka Dončić**

2) **Shai Gilgeous-Alexander**

3) **Giannis Antetokounmpo**

4) **Nikola Jokic**

5) **Victor Wembanyama**

These players all demonstrate elite all-around stat lines, showing the model's emphasis on versatility and high usage. While the model doesn't account for team success or media narratives, it effectively highlights players with the most MVP-worthy statistical profiles.

Conclusion:

This project explored the evolution of NBA player performance. I did this using statistical methods, such as Anova Tables, Gini Coefficients, and several Regression Models, and a comprehensive dataset spanning nearly 70 seasons, from the 1950s all the way till 2024. By analyzing specific counting statistical trends over time and factors influencing individual and career success, I was able to find several meaningful patterns in how the games play has changed.

Our decade-based trend analysis helped in revealing the NBA's shift from a -man, post-oriented game to a faster-paced, perimeter-focused, guard dominated style. Points per game have fluctuated with changes in tempo and rules, meanwhile, assists and three-point play have risen significantly in the recent modern era. This is reflective of more team-oriented and analytics-driven strategies.

When I examined performance by position, clear statistical trends emerged. Guards dominate in assists and steals, centers dominate in rebounds and blocks, and forwards maintain balance between the two. These differences were confirmed through ANOVA testing, with post-hoc comparisons highlighting significant gaps between roles across all the major stat categories.

The Gini coefficient analysis showed that scoring inequality was at its highest in earlier decades, dipped in the 1970s, and later rose again with the emergence of high usage player-centric offenses. More recently, scoring has become increasingly balanced, due to more distributed offensive systems and deeper team rotations.

For the outlier analysis, I used combined Z-scores to identify the most statistically dominant individual seasons in NBA history. Wilt Chamberlain, Oscar Robertson, Russell Westbrook, and Nikola Jokic stood out as players with multiple historically great seasons, which was defined by across-the-board contributions in scoring, rebounding, and passing.

When predicting career length, the regression model I made found that players who consistently contribute in multiple areas, such as assists and rebounds, tended to have longer NBA careers. This suggests that all around impact may be more important to longevity than pure scoring ability.

Finally, I used logistic regression to predict MVP results. The model demonstrated strong performance, correctly highlighting top 2024 MVP contenders like Luka Dončić, Shai Gilgeous-Alexander, Giannis Antetokounmpo, and the winner of the award that year, Nikola Jokic. It showed that high production across points, assists, rebounds, and minutes are essential to being an MVP caliber player.

Overall, this analysis confirms that the game of basketball, and the statistical trends within the game, has continually evolved. While raw scoring remains important, long-term success and MVP status increasingly favor versatility, efficiency, and adaptability. Statistical tools not only helped us in understanding past greatness but also offered us predictive power to evaluate current and future stars

Appendix

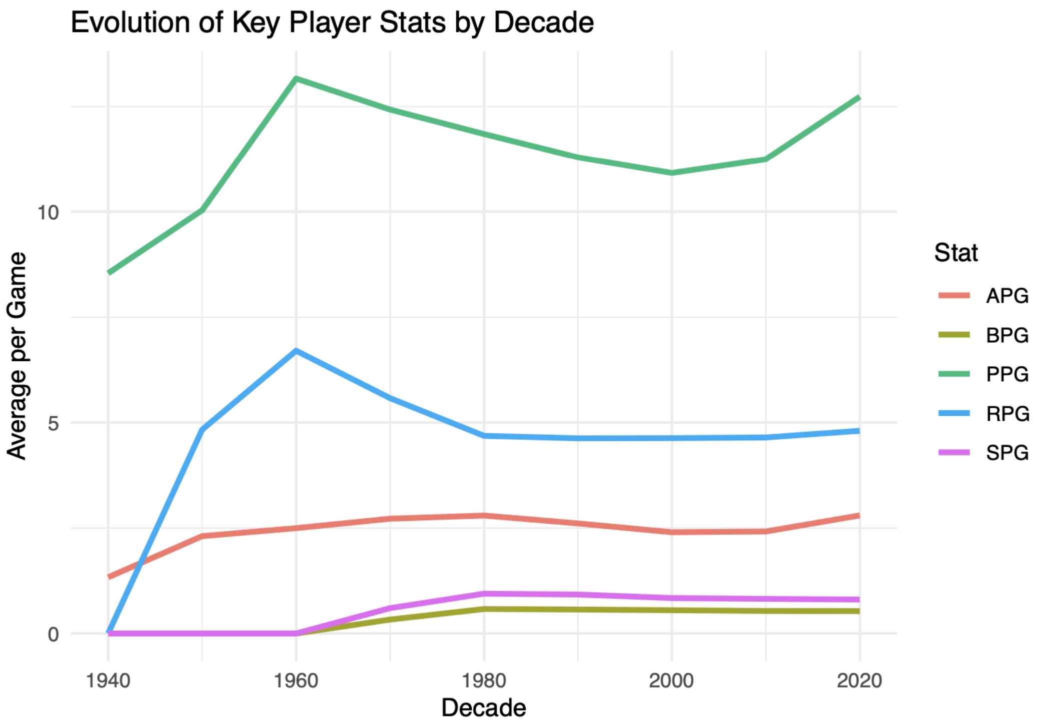
Code 1

```
knitr::opts_chunk$set(echo = TRUE, message = FALSE, warning = FALSE)

library(tidyverse)

library(ggplot2)

library(lubridate)

library(GGally)

library(broom)

library(readxl)

library(ineq)

nba <- read_excel("Player_Per_Game_Filtered.xlsx", skip = 1)

names(nba) <- str_trim(names(nba))

names(nba) <- tolower(names(nba))
```

Code 2

```
nba$decade <- floor(nba$season / 10) * 10

stat_trends <- nba %>%

 group_by(decade) %>%

  summarize(PPG = mean(pts_per_game, na.rm = TRUE),

  RPG = mean(trb_per_game, na.rm = TRUE),

  APG = mean(ast_per_game, na.rm = TRUE),

  SPG = mean(stl_per_game, na.rm = TRUE),

  BPG = mean(blk_per_game, na.rm = TRUE),

  .groups = "drop")

stat_trends_long <- stat_trends %>% pivot_longer(cols = -decade, names_to = "Stat",

values_to = "Value")

ggplot(stat_trends_long, aes(x = decade,
y = Value, color = Stat)) + geom_line(size = 1.2) +

labs(title = "Evolution of Key Player Stats by Decade",

x = "Decade", y = "Average per Game") + theme_minimal()
```

Figure 1

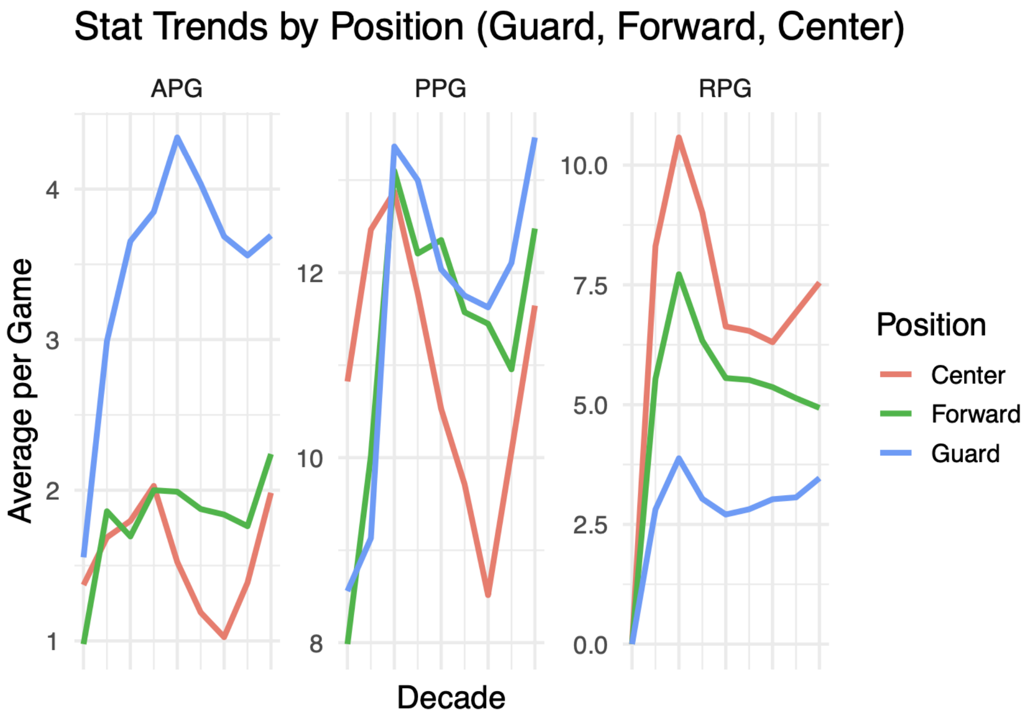


Code 3

```
nba <- nba %>%
  mutate(primary_pos = case_when(
   str_detect(pos, "G") ~ "Guard",
    str_detect(pos, "F") ~ "Forward",
str_detect(pos, "C") ~ "Center",
   TRUE ~ NA_character_))
nba_filtered <- nba %>%
  filter(
 !is.na(season),
    !is.na(primary_pos),
    !is.na(pts_per_game),
    !is.na(ast_per_game),
    !is.na(trb_per_game)
) %>%
  mutate(decade = floor(as.numeric(season) / 10) * 10)
position_trends <- nba_filtered %>%
  group_by(decade, primary_pos) %>%
  summarize(
   PPG = mean(pts_per_game, na.rm = TRUE),
    RPG = mean(trb_per_game, na.rm = TRUE),
    APG = mean(ast_per_game, na.rm = TRUE),
    .groups = "drop")
position_trends_long <- position_trends %>%
  pivot_longer(
   cols = c(PPG, RPG, APG),
    names_to = "Stat",
    values_to = "Value")
```

```
ggplot(position_trends_long, aes(x = decade, y = Value, color = primary_pos)) +

  geom_line(size = 1.2) +

  facet_wrap(~ Stat, scales = "free_y") +

  labs(title = "Stat Trends by Position (Guard, Forward, Center)", x = "Decade", # keep label
y = "Average per Game",
color = "Position")+
theme_minimal(base_size = 14) + theme(

   axis.text.x = element_blank()  )
```

Figure 2

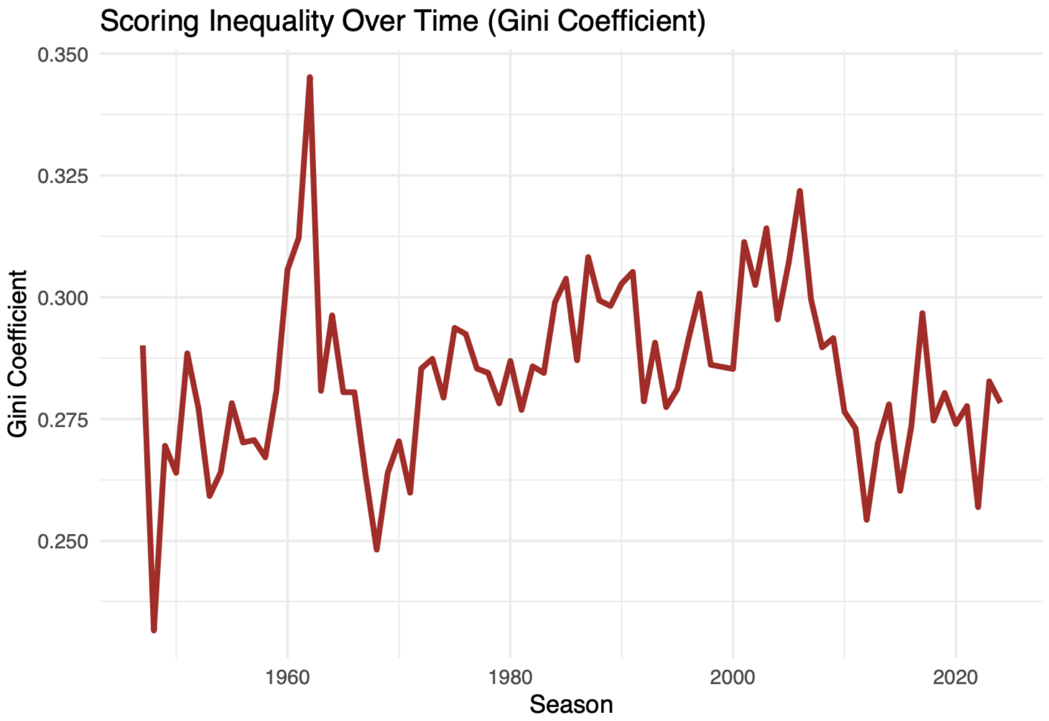


Code 4

```
gini_trend <- nba %>%

  group_by(season) %>%

  summarize(

   gini_ppg = ineq(pts_per_game, type = "Gini", na.rm = TRUE))
```

```
ggplot(gini_trend, aes(x = season, y = gini_ppg)) +
  geom_line(color = "firebrick", size = 1.2) +
  labs(
title = "Scoring Inequality Over Time (Gini Coefficient)",
    x = "Season",
    y = "Gini Coefficient")+ theme_minimal()
```

Figure 3

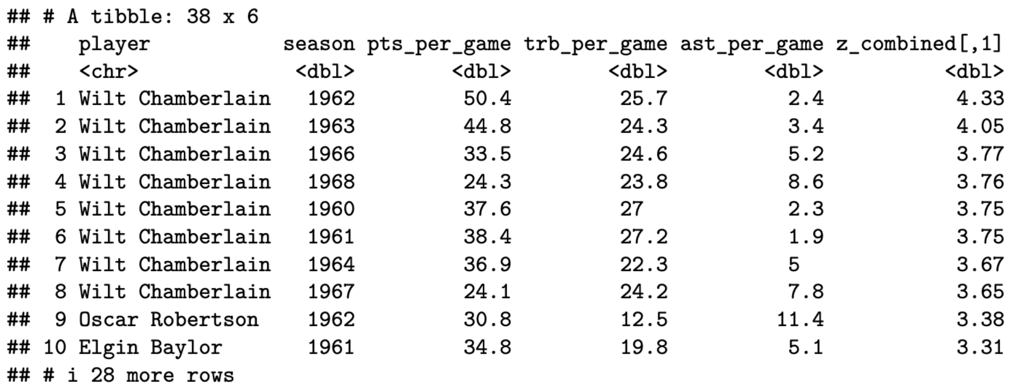


" " " Interpretation

Code 5

```
nba_outliers <- nba %>%
  mutate(
   z_ppg = scale(pts_per_game),
    z_rpg = scale(trb_per_game),

    z_apg = scale(ast_per_game),

    z_combined = (z_ppg + z_rpg + z_apg) / 3)

top_players_combined <- nba_outliers %>%

  filter(z_combined > 2.575) %>%

  select(

player, season, pts_per_game,

trb_per_game, ast_per_game, z_combined

  ) %>%

arrange(desc(z_combined))

top_players_combined
```

Figure 4

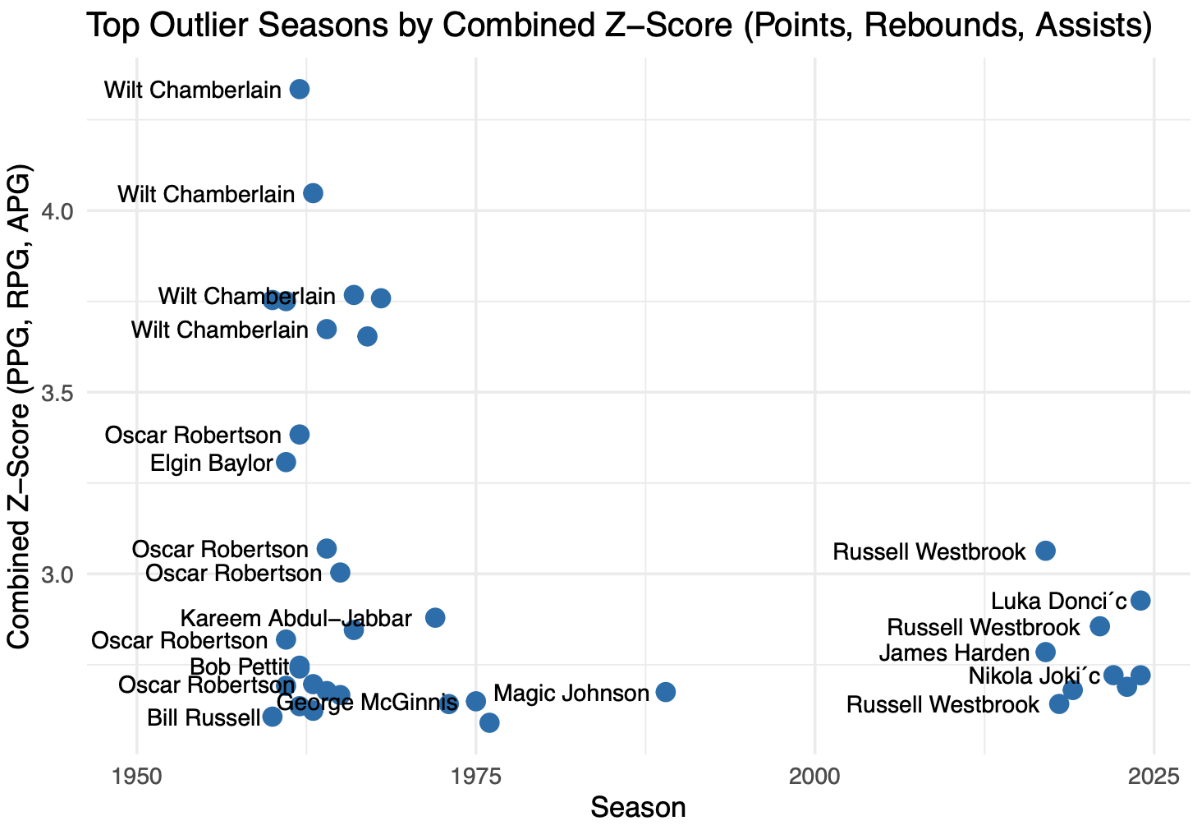


```
## # A tibble: 38 x 6
##    player            season pts_per_game trb_per_game ast_per_game z_combined[,1]
##    <chr>              <dbl>        <dbl>        <dbl>        <dbl>          <dbl>
##  1 Wilt Chamberlain   1962         50.4         25.7          2.4           4.33
##  2 Wilt Chamberlain   1963         44.8         24.3          3.4           4.05
##  3 Wilt Chamberlain   1966         33.5         24.6          5.2           3.77
##  4 Wilt Chamberlain   1968         24.3         23.8          8.6           3.76
##  5 Wilt Chamberlain   1960         37.6         27            2.3           3.75
##  6 Wilt Chamberlain   1961         38.4         27.2          1.9           3.75
##  7 Wilt Chamberlain   1964         36.9         22.3          5             3.67
##  8 Wilt Chamberlain   1967         24.1         24.2          7.8           3.65
##  9 Oscar Robertson    1962         30.8         12.5         11.4           3.38
## 10 Elgin Baylor       1961         34.8         19.8          5.1           3.31
## # i 28 more rows
```

Code 6

```
ggplot(top_players_combined, aes(x = season, y = z_combined, label = player)) +

  geom_point(color = "#0072B2", size = 3) +

  geom_text(hjust = 1.1, size = 3, check_overlap = TRUE) +

  labs(

   title = "Top Outlier Seasons by Combined Z-Score (Points, Rebounds, Assists)",

    x = "Season",
    y = "Combined Z-Score (PPG, RPG, APG)")+
scale_x_continuous(limits = c(1950, NA)) +
theme_minimal()
```

Figure 5



Code 7

```
career_data <- nba %>%
  group_by(player) %>%
  summarize(
   career_length = n(),
    avg_ppg = mean(pts_per_game, na.rm = TRUE),
    avg_rpg = mean(trb_per_game, na.rm = TRUE),
    avg_apg = mean(ast_per_game, na.rm = TRUE),
  .groups = "drop")

career_model <- lm(career_length ~ avg_ppg + avg_rpg + avg_apg, data = career_data)

summary(career_model)
```

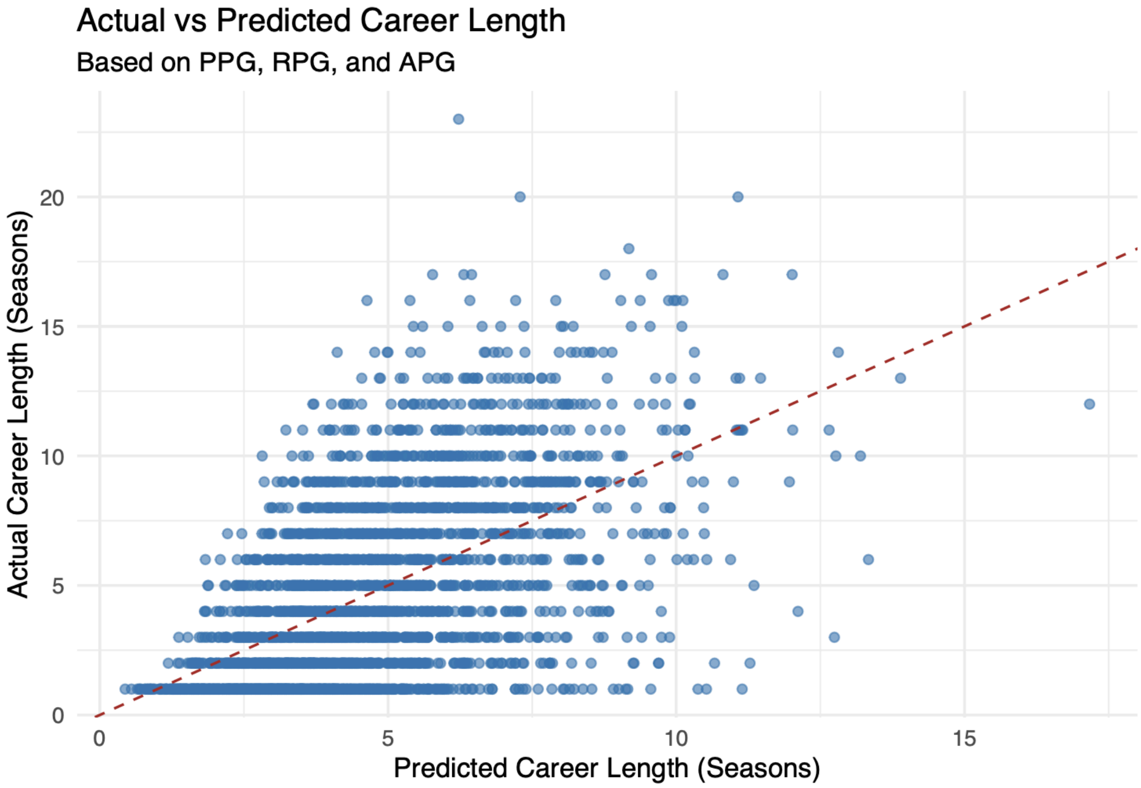
Figure 6

```
##
## Call:
## lm(formula = career_length ~ avg_ppg + avg_rpg + avg_apg, data = career_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.1444  -1.8881  -0.4638   1.6810  16.7766
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.09816    0.13419   0.732    0.465
## avg_ppg      0.19284    0.01642  11.741   <2e-16 ***
## avg_rpg      0.38975    0.02553  15.264   <2e-16 ***
## avg_apg      0.51837    0.04479  11.574   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.938 on 2844 degrees of freedom
## Multiple R-squared:  0.3285, Adjusted R-squared:  0.3278
## F-statistic: 463.8 on 3 and 2844 DF,  p-value: < 2.2e-16
```

Code 8

```
career_data$predicted_length <- predict(career_model)

ggplot(career_data, aes(x = predicted_length, y = career_length)) +

 geom_point(alpha = 0.6, color = "#1f77b4") +

 geom_abline(slope = 1, intercept = 0, linetype = "dashed", color = "firebrick") +

 labs(

  title = "Actual vs Predicted Career Length",

   subtitle = "Based on PPG, RPG, and APG",

   x = "Predicted Career Length (Seasons)",

   y = "Actual Career Length (Seasons)"

9)+ theme_minimal()
```

Figure 7



Code 9

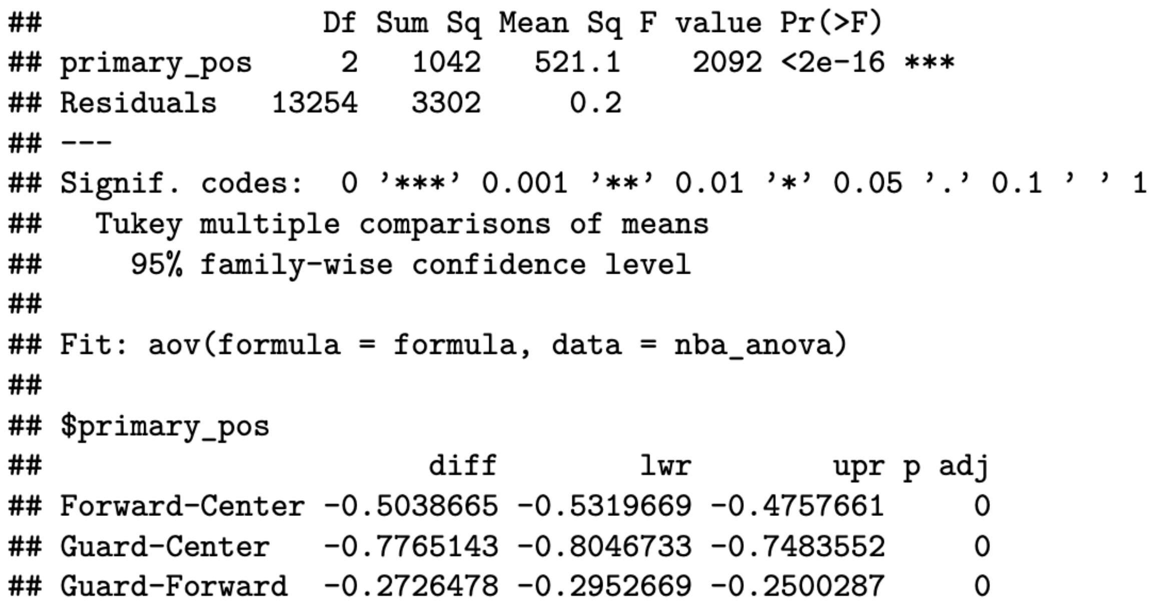
```
nba_anova <- nba %>%

  mutate(primary_pos = case_when(

   str_detect(pos, "G") ~ "Guard",

    str_detect(pos, "F") ~ "Forward",

    str_detect(pos, "C") ~ "Center",

    TRUE ~ NA_character_

 )) %>%

  filter(!is.na(primary_pos))

run_anova <- function(stat) {
formula <- as.formula(paste(stat, "~ primary_pos")) result <- aov(formula, data = nba_anova)
print(summary(result))
print(TukeyHSD(result))

}
```

```
aov_ppg <- run_anova("pts_per_game")

aov_apg <- run_anova("ast_per_game")

aov_rpg <- run_anova("trb_per_game")

aov_spg <- run_anova("stl_per_game")

aov_bpg <- run_anova("blk_per_game")
```

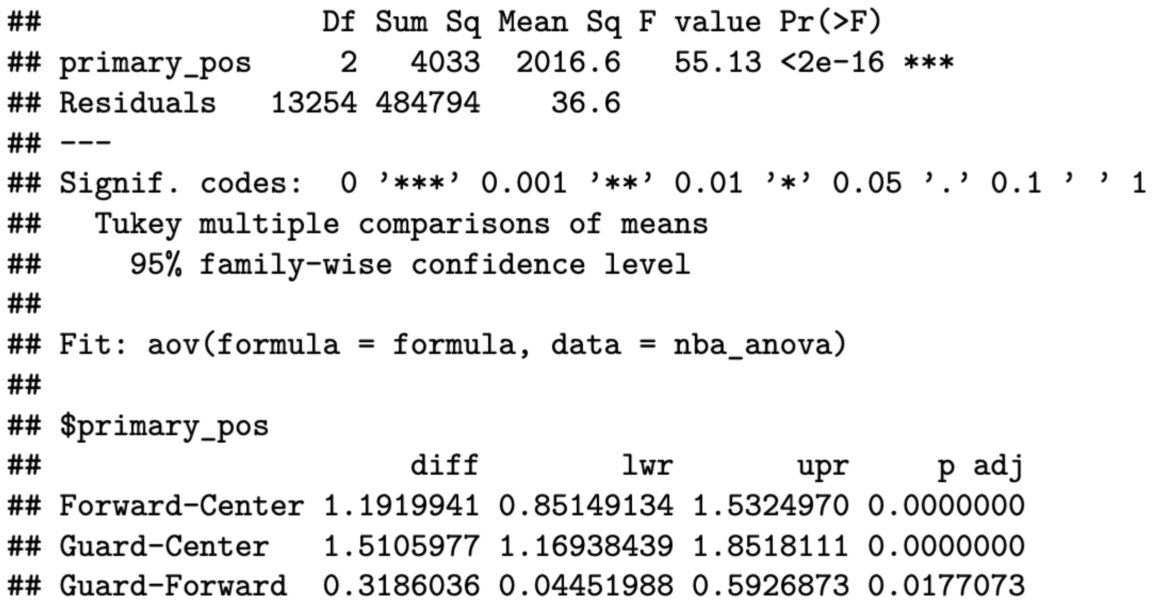
Figure 8

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## primary_pos     2  12764    6382    2251 <2e-16 ***
## Residuals   13254  37573       3
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = formula, data = nba_anova)
##
## $primary_pos
##                     diff       lwr       upr p adj
## Forward-Center 0.3889481 0.2941541 0.4837422     0
## Guard-Center   2.2452193 2.1502275 2.3402112     0
## Guard-Forward  1.8562712 1.7799679 1.9325746     0
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## primary_pos     2  37854   18927    2773 <2e-16 ***
## Residuals   13254  90451       7
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = formula, data = nba_anova)
##
## $primary_pos
##                     diff       lwr       upr p adj
## Forward-Center -1.732042 -1.879121 -1.584964     0
## Guard-Center   -4.376848 -4.524234 -4.229463     0
## Guard-Forward  -2.644806 -2.763195 -2.526417     0
##                Df Sum Sq Mean Sq F value Pr(>F)
## primary_pos     2    237   118.4   399.9 <2e-16 ***
## Residuals   13254   3924     0.3
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = formula, data = nba_anova)
##
## $primary_pos
##                     diff       lwr       upr p adj
## Forward-Center 0.1643308 0.1336954 0.1949663     0
## Guard-Center   0.3557111 0.3250117 0.3864105     0
## Guard-Forward  0.1913802 0.1667206 0.2160399     0
```



```
##                 Df Sum Sq Mean Sq F value Pr(>F)
## primary_pos      2   1042   521.1    2092 <2e-16 ***
## Residuals    13254   3302     0.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = formula, data = nba_anova)
##
## $primary_pos
##                      diff        lwr        upr p adj
## Forward-Center -0.5038665 -0.5319669 -0.4757661     0
## Guard-Center   -0.7765143 -0.8046733 -0.7483552     0
## Guard-Forward  -0.2726478 -0.2952669 -0.2500287     0
```

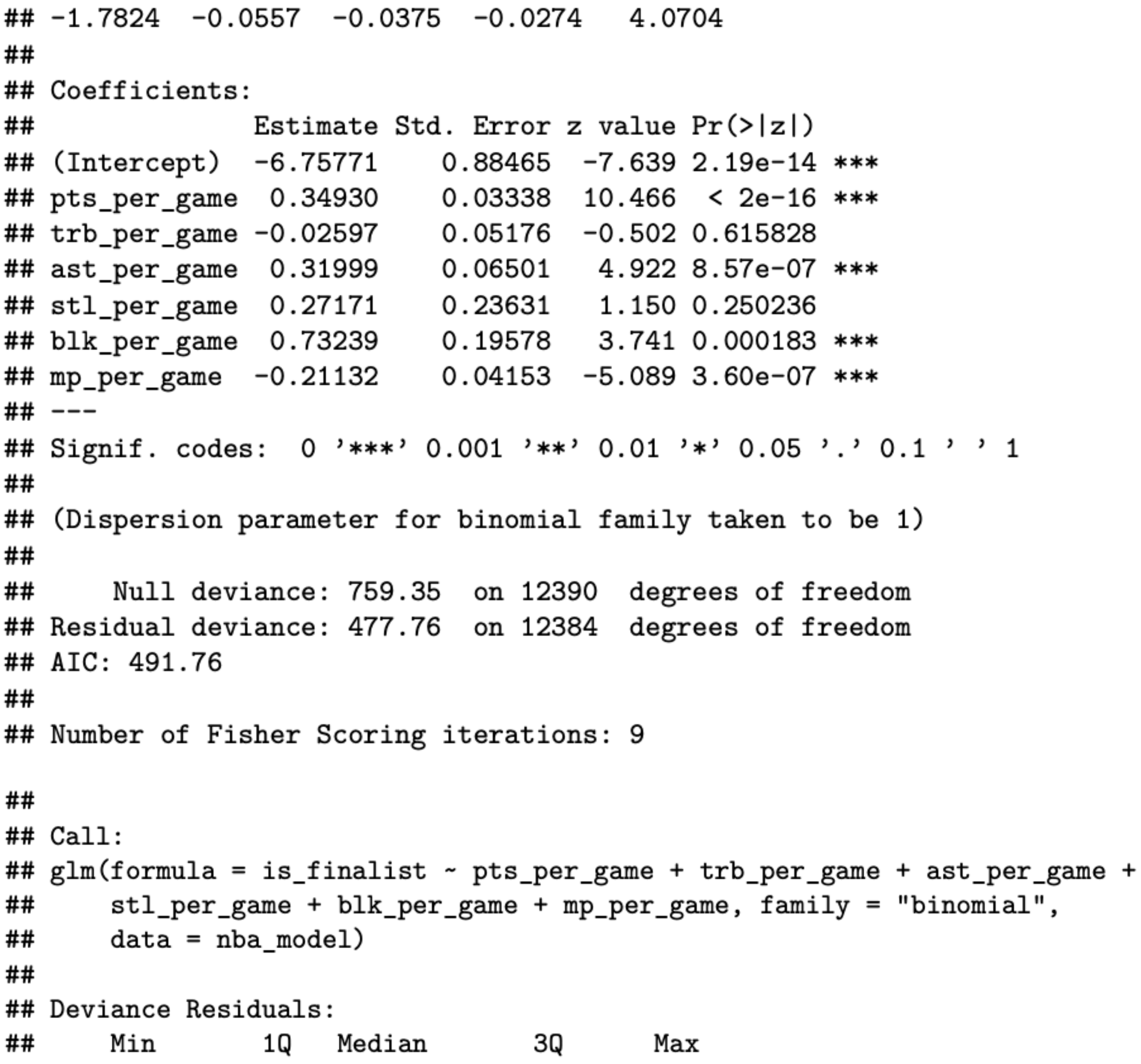


```
##                 Df Sum Sq Mean Sq F value Pr(>F)
## primary_pos      2   4033  2016.6   55.13 <2e-16 ***
## Residuals    13254 484794    36.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = formula, data = nba_anova)
##
## $primary_pos
##                     diff        lwr       upr     p adj
## Forward-Center 1.1919941 0.85149134 1.5324970 0.0000000
## Guard-Center   1.5105977 1.16938439 1.8518111 0.0000000
## Guard-Forward  0.3186036 0.04451988 0.5926873 0.0177073
```

Code 10

```
finalists <- c(
  "Nikola Jokic", "Joel Embiid", "Giannis Antetokounmpo",
  "Luka Doncic", "Jayson Tatum", "Stephen Curry",
  "Kevin Durant", "LeBron James", "Shai Gilgeous-Alexander")
nba_model <- nba %>%
  filter(season >= 1955, season <= 2023) %>%
  mutate(is_finalist = ifelse(player %in% finalists, 1, 0)) %>%
  select(
   is_finalist, pts_per_game, trb_per_game,
   ast_per_game, stl_per_game, blk_per_game, mp_per_game
  ) %>%
drop_na()
model <- glm(
  is_finalist ~ pts_per_game + trb_per_game + ast_per_game +
   stl_per_game + blk_per_game + mp_per_game,
  data = nba_model,
  family = "binomial"
)
summary(model)
```

Figure 9



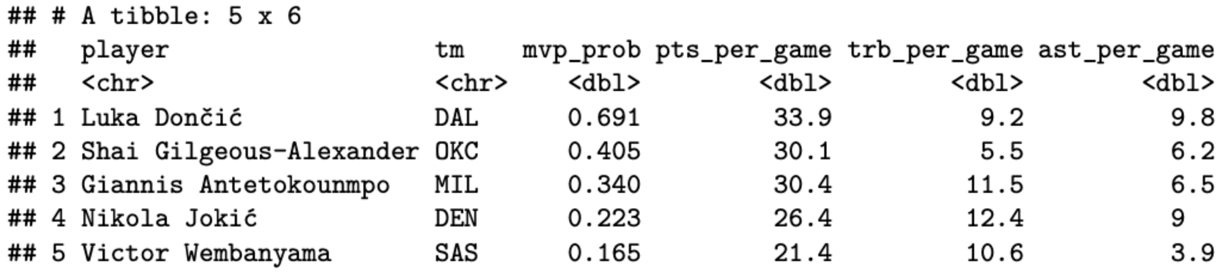
```
## -1.7824  -0.0557  -0.0375  -0.0274   4.0704
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -6.75771    0.88465  -7.639 2.19e-14 ***
## pts_per_game  0.34930    0.03338  10.466  < 2e-16 ***
## trb_per_game -0.02597    0.05176  -0.502 0.615828
## ast_per_game  0.31999    0.06501   4.922 8.57e-07 ***
## stl_per_game  0.27171    0.23631   1.150 0.250236
## blk_per_game  0.73239    0.19578   3.741 0.000183 ***
## mp_per_game  -0.21132    0.04153  -5.089 3.60e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 759.35  on 12390  degrees of freedom
## Residual deviance: 477.76  on 12384  degrees of freedom
## AIC: 491.76
##
## Number of Fisher Scoring iterations: 9

##
## Call:
## glm(formula = is_finalist ~ pts_per_game + trb_per_game + ast_per_game +
##     stl_per_game + blk_per_game + mp_per_game, family = "binomial",
##     data = nba_model)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

Code 11

```
nba_2024 <- nba %>%
 filter(season == 2024) %>%
 drop_na(
  pts_per_game, trb_per_game, ast_per_game,
  stl_per_game, blk_per_game, mp_per_game
 )
```

```
nba_2024$mvp_prob <- predict(model, newdata = nba_2024, type = "response")

top_5 <- nba_2024 %>%

  arrange(desc(mvp_prob)) %>%

  select(

   player, tm, mvp_prob,

   pts_per_game, trb_per_game, ast_per_game

  ) %>%

 slice_head(n = 5)

top_5
```

Figure 10



```
## # A tibble: 5 x 6
##   player                  tm    mvp_prob pts_per_game trb_per_game ast_per_game
##   <chr>                   <chr>    <dbl>        <dbl>        <dbl>        <dbl>
## 1 Luka Dončić             DAL      0.691         33.9          9.2          9.8
## 2 Shai Gilgeous-Alexander OKC      0.405         30.1          5.5          6.2
## 3 Giannis Antetokounmpo   MIL      0.340         30.4         11.5          6.5
## 4 Nikola Jokić            DEN      0.223         26.4         12.4          9  
## 5 Victor Wembanyama       SAS      0.165         21.4         10.6          3.9
```