Abdullah Al Iman
City College of New York
Econ B2000
Prof. Kevin Foster
Dec. 2024

Hot Hand Analysis

Introduction:

When watching sports, especially Basketball, it is very common to hear the phrase “He’s on Fire!” and “[Player’s name] has the hot hand tonight!”. This “Hot Hand” refers to the ability of making a certain amount of shots consecutively. In basic terms, if a player makes at least one shot, his chances of making the next shot increases. This is a phenomenon that has been discussed throughout the NBA’s existence. Does the previous shot’s result increase the chances of making the next shot? This is the topic of the data and code that will be explored.

Before delving into the data, the null hypothesis will be that the result of the previous shot will have no effect on the next. The alternative will be that the result of the previous shot will have an effect. The level of significance we will use is 0.05. This means that if the p-value is greater than 0.05, we can conclude from the data that the result of the previous has no effect on the next. The techniques that will be utilized in the data are logistic regression, random forest, visualizations and a monte carlo simulation.

I also want to discuss articles that I had read prior to looking into dara to get a brief overview on past research and discourse on the “Hot Hand” discourse. The academic articles that I focused on were “The Effect of the Three-Point Line on Basketball Scoring and Strategy” by R.D. Hill and J.J. Pritchard and “Scoring, Efficiency, and Rule Changes in the NBA: The

Abdullah Al Iman
City College of New York
Econ B2000
Prof. Kevin Foster
Dec. 2024

Evolution of Basketball” by C.H. Stone and A.S. Evans. These articles help me do further research on how the NBA’s scoring average has varied throughout the years.

For “The Effect of the Three-Point Line on Basketball Scoring and Strategy” , it analyzes the implementation of the 3 point line and how that impacted scoring trends afterwards. It also includes how teams and players have spaced out and how that impacted the defensive side as well. The changes in strategies that coaches have integrated were also highlighted.

For “Scoring, Efficiency, and Rule Changes in the NBA: The Evolution of Basketball” , it analyzes the impact of rule changes on scoring. Rule changes include shot clock violations, hand checking and defensive rule variation. They mention the scoring efficiency as well and how it had long term and short term effects when those rules were adjusted.

Econometrics techniques that can be used when utilizing these articles for my project would be regressions and models. They would visualize the trends and relationships between rule changes/new implementations and scoring over the decades.

Other articles that added further research and depth to my topic are “The Hot Hand in Basketball" by Gilovich, Vallone, & Tversky and "Hot Hand Fallacy: Cognitive Mistakes or Equilibrium Adjustments?" by Yigal Attali.

In “The Hot Hand in Basketball” , the main point is debating whether or not the “hot hand” has truth to it. It explores the concept of the success rate of a player’s shot attempt if that player had already been making a certain number of shots in a row. It was concluded that it was just a “random sequence” rather than having a “hot hand”. An Econometric technique that can be used based on the data and topic is the theory of probability.

Abdullah Al Iman
City College of New York
Econ B2000
Prof. Kevin Foster
Dec. 2024

In “Hot Hand Fallacy: Cognitive Mistakes or Equilibrium Adjustments” , the author explores the same concept of whether or not a “hot hand” truly exists. The author asks the question whether it is the player’s skill or if it is the strategy of the team. He came to the conclusion that once a player makes multiple shots in a row, the player’s shot style and difficulty changes and that has an impact on their percentages. Econometric techniques used based on this article include regressions and time series analysis.

Code/Data

```
library(dplyr)
library(purrr)
library(hoopR)
library(lme4)
library(broom)
library(ggplot2)
library(randomForest)
library(caret)
```

The first part of the data/code is inputting the necessary packages. The first package is DPLYR which is required for manipulating data and filtering it. This will allow us to group shots by players and calculate their shooting success. “Purrr” allows us to get shot data for observed players and seasons in one step, It essentially helps us gather a list of data much quicker. The “hoopR” library gives access to NBA data such as the players’ shot chart and the seasons that they played. It gives us all the information we need without manually collecting data. The “lme4” library helps us create statistical models such as logistic regression as we will be

Abdullah Al Iman
City College of New York
Econ B2000
Prof. Kevin Foster
Dec. 2024
observing a binary dependent variable. The “broom” library will clean up the regression gained from the “lme4” library. “Ggplot2” will give us the ability to create graphs and bring the data to life by inputting visuals such as histograms, bar graphs and line graphs. “Randomforest” will help determine which factor will be the most vital when it comes to figuring out whether a shot goes in or not. “Caret” will split the data into testing and training sets.

```
players <- list(
  "Steph" = list(id = 201939, seasons = c("2009-10", "2010-11", "2011-12", "2012-13"
  "Harden" = list(id = 201935, seasons = c("2009-10", "2010-11", "2011-12", "2012-13
  "LeBron" = list(id = 2544, seasons = c("2003-04", "2004-05", "2005-06", "2006-07",
  "Durant" = list(id = 201142, seasons = c("2007-08", "2008-09", "2009-10", "2010-11
  "Giannis" = list(id = 203507, seasons = c("2013-14", "2014-15", "2015-16", "2016-1
  "Klay" = list(id = 202691, seasons = c("2011-12", "2012-13", "2013-14", "2014-15",
  "Westbrook" = list(id = 201566, seasons = c("2008-09", "2009-10", "2010-11", "2011
  "Kyrie" = list(id = 202681, seasons = c("2011-12", "2012-13", "2013-14", "2014-15"
  "Damian" = list(id = 203081, seasons = c("2012-13", "2013-14", "2014-15", "2015-16
  "Paul George" = list(id = 202331, seasons = c("2010-11", "2011-12", "2012-13", "20
)
```

The next section of the code is gathering the players that will be observed as well as the seasons that they have played since getting drafted. This information was gathered from the “hoopR” package. The players are identified through their player id (E.G. Stephen Curry is identified through “id = 201939”).

Abdullah Al Iman
City College of New York
Econ B2000
Prof. Kevin Foster
Dec. 2024

```
all_shots <- purrr::map_df(names(players), function(player_name) {
  player_info <- players[[player_name]]
  player_id <- player_info$id
  player_seasons <- player_info$seasons

  purrr::map_df(player_seasons, function(season) {
    res <- tryCatch({
      nba_shotchartdetail(player_id = player_id, season = season)
    }, error = function(e) {
      message(paste("Error for", player_name, "in", season, ":", e$message))
      return(NULL)
    })

    if (!is.null(res) && "Shot_Chart_Detail" %in% names(res)) {
      res$Shot_Chart_Detail %>%
        mutate(
          id = player_id,
          name = player_name,
          season = season
        )
    } else {
      return(tibble())
    }
  })
})
```

This code is a loop that lets us gather the shot data of the observed players in the specified seasons that player played in. We are using the “nba_shotchartdetail” function to do this as this was gained from the “hoopR” package. To handle issues such as injuries or suspensions, the “tryCatch” function will be used.

```
shots_clean <- all_shots %>%
  mutate(
    made = as.numeric(SHOT_MADE_FLAG),
    prev_shot = lag(as.numeric(SHOT_MADE_FLAG), 1),
    distance = as.numeric(SHOT_DISTANCE),
    is_2pt = ifelse(SHOT_TYPE == "2PT Field Goal", 1, 0),
    move_type = as.factor(ACTION_TYPE),
    zone = as.factor(SHOT_ZONE_BASIC),
    is_home = ifelse(TEAM_NAME == HTM, 1, 0),
    name = as.factor(name)
  ) %>%
  arrange(name, GAME_ID, PERIOD, desc(MINUTES_REMAINING), desc(SECONDS_REMAINING)) %>%
  group_by(name, GAME_ID) %>%
  mutate(prev_shot = lag(made, 1)) %>%
  ungroup() %>%
  filter(!is.na(prev_shot)) %>%
  select(made, prev_shot, distance, is_2pt, move_type, zone, is_home, name)
```

Abdullah Al Iman
City College of New York
Econ B2000
Prof. Kevin Foster
Dec. 2024

This part filters and cleans the data that was found from the previous section. The shot made variable is introduced and is labeled as “made” and variables such as “prev_shot” and “distance” are also created. These variables represent the result of the previous shot and the distance of that shot respectively.

```
hot_hand_model <- glmer(
  made ~ prev_shot + distance + is_2pt + zone + is_home + move_type + (1 | name),
  data = shots_clean,
  family = binomial(link = "logit"),
  nAGQ = 0,
  control = glmerControl(optimizer = "nloptwrap")
)

summary(hot_hand_model)
```

This is where we attain the logistic regression model to actually test whether the “Hot Hand” is real or not based on the data. As a reminder, we are trying to observe if the result of the previous shot which is denoted as “prev_shot” affects the chances of the next shot, denoted as “made”. Factors such as distance, whether the player was home or away, the location of the shot were all taken into consideration as well.

Abdullah Al Iman
City College of New York
Econ B2000
Prof. Kevin Foster
Dec. 2024

```
simulations <- 10
rand_coefs <- numeric(simulations)

set.seed(123)
for (i in seq_len(simulations)) {
  rand_data <- shots_clean %>%
    mutate(prev_shot = sample(prev_shot))

  rand_model <- glmer(
    made ~ prev_shot + distance + is_2pt + zone + is_home + move_type + (1 | name),
    data = rand_data,
    family = binomial(link = "logit"),
    nAGQ = 0,
    control = glmerControl(optimizer = "nloptwrap")
  )

  rand_coefs[i] <- summary(rand_model)$coefficients["prev_shot", "Estimate"]
}


actual_coef <- summary(hot_hand_model)$coefficients["prev_shot", "Estimate"]
cat("Actual coefficient for prev_shot:", actual_coef, "\n")
cat("Mean of randomized coefficients:", mean(rand_coefs), "\n")


ggplot(data.frame(rand_coefs), aes(x = rand_coefs)) +
  geom_histogram(binwidth = 0.01, fill = "blue", alpha = 0.7) +
  geom_vline(xintercept = actual_coef, color = "red", linetype = "dashed", size = 1) +
  labs(
    title = "Monte Carlo Simulation: Coefficients for Previous Shot",
    x = "Randomized Coefficients",
    y = "Frequency"
  ) +
  theme_minimal()
```

The Monte Carlo Simulation will be used to see the actual results and not just predicted probability from the logistic regression model. It will see if the “Hot Hand” is random or not. The histogram will showcase it.

Abdullah Al Iman
City College of New York
Econ B2000
Prof. Kevin Foster
Dec. 2024

```
split <- createDataPartition(shots_clean$made, p = 0.7, list = FALSE)
train <- shots_clean[split, ]
test <- shots_clean[-split, ]

rf_model <- randomForest(
  made ~ prev_shot + distance + is_2pt + move_type + zone + is_home + name,
  data = train,
  ntree = 500,
  importance = TRUE
)


varImpPlot(rf_model)
```

The Random Forest model will find information on which factor impacts the results of the “Hot Hand” analysis the most such as shot distance and zone.

```
success_rate <- shots_clean %>%
  group_by(name, GAME_ID) %>%
  summarize(success_rate = mean(made, na.rm = TRUE), .groups = "drop")

ggplot(success_rate, aes(x = GAME_ID, y = success_rate, color = name, group = name)) +
  geom_line() +
  labs(
    title = "Success Rate Over Games",
    x = "Game ID",
    y = "Success Rate",
    color = "Player"
  ) +
  theme_minimal()


zone_count <- shots_clean %>%
  group_by(name, zone) %>%
  summarize(total_shots = n(), .groups = "drop")

ggplot(zone_count, aes(x = zone, y = total_shots, fill = name)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(
    title = "Shot Distribution by Zone",
    x = "Shot Zone",
    y = "Total Shots",
    fill = "Player"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

ggplot(success_rate, aes(x = success_rate, fill = name)) +
  geom_histogram(binwidth = 0.05, alpha = 0.7, position = "dodge") +
  labs(
    title = "Distribution of Success Rates",
```

These graphs give a visual representation of the shot chart of each observed player.

Results:

Abdullah Al Iman
City College of New York
Econ B2000
Prof. Kevin Foster
Dec. 2024

This next section will now showcase the results of the code/data. This will include the results of the logistic regression , the monte carlo simulation and the results of the graphs. There will also be discussion on why the results might be the case.

For the Logistic Regression we get the results :

```
Generalized linear mixed model fit by maximum likelihood (Adaptive Gauss-Hermite
  Quadrature, nAGQ = 0) [glmerMod]
 Family: binomial  ( logit )
Formula: made ~ prev_shot + distance + is_2pt + zone + is_home + move_type +
    (1 | name)
   Data: shots_clean
Control: glmerControl(optimizer = "nloptwrap")

     AIC      BIC   logLik deviance df.resid
189970.9 190754.6 -94906.5 189812.9   150253

Scaled residuals:
    Min      1Q  Median      3Q     Max
-7.6391 -0.8210 -0.6168  0.9984  5.3000

Random effects:
 Groups Name        Variance Std.Dev.
 name   (Intercept) 0.02306  0.1518
Number of obs: 150332, groups:  name, 10

Fixed effects:
                                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                               2.320668   0.171038  13.568  < 2e-16 ***
prev_shot                                -0.036122   0.011044  -3.271 0.001073 **
distance                                 -0.003835   0.002536  -1.512 0.130459
is_2pt                                    0.328461   0.370226   0.887 0.374977
zoneBackcourt                            -2.472599   0.305520  -8.093 5.82e-16 ***
zoneIn The Paint (Non-RA)                -0.608630   0.371928  -1.636 0.101751
zoneLeft Corner 3                         0.259681   0.038966   6.664 2.66e-11 ***
zoneMid-Range                            -0.327217   0.370335  -0.884 0.376929
zoneRestricted Area                       0.057929   0.374154   0.155 0.876957
zoneRight Corner 3                        0.284702   0.038017   7.489 6.95e-14 ***
```

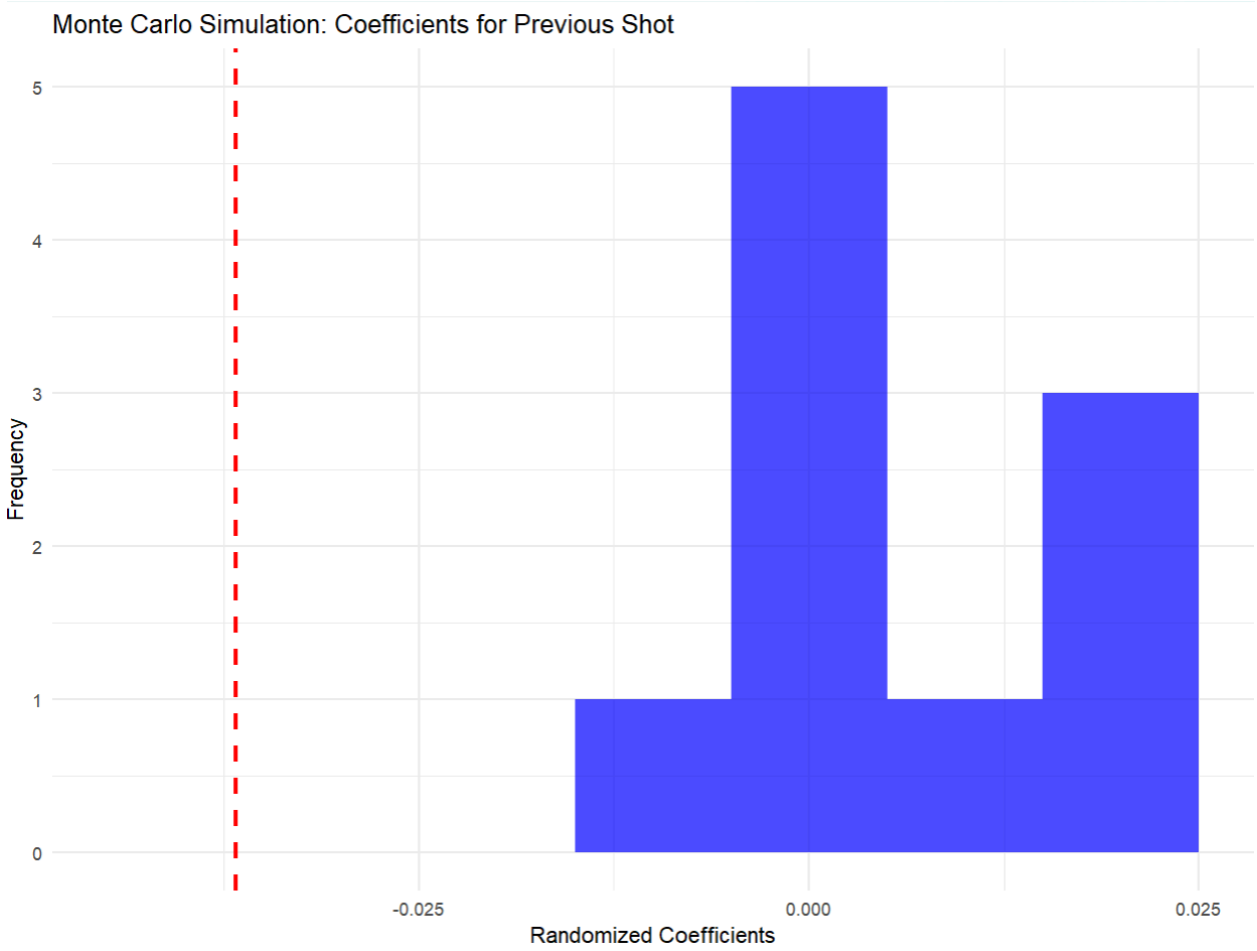
Based on our data, which includes over 150,000 shots from the 10 players we observed we saw the “impact” of the variables. For the previous shot, the coefficient is negative which means that making the previous shot actually decreases the chances of making the next. The p-value is 0.001 , making it statistically significant which means that based on the data, this is

Abdullah Al Iman
City College of New York
Econ B2000
Prof. Kevin Foster
Dec. 2024
not random. This is a very interesting result as it completely goes against what the “Hot Hand” is. When we look at distance , we also see a negative coefficient which means as the distance increases , the chances of making a shot decreases. This makes more sense in the context of Basketball. This also goes in hand with the results from the “is_2pt” variable as it has a positive coefficient , which suggests that a 2 point shot has a higher chance of going in than a 3 point shot. We also have other results that go hand in hand with basketball context as the “cutting dunk shots” and “slam dunk shots” have positive coefficients while “Fadeaway Jump Shots” have less likely to succeed. These make sense as dunks are an easier shot type and will have a higher percentage of going in.

```
> cat("Actual coefficient for prev_shot:", actual_coef, "\n")
Actual coefficient for prev_shot: -0.03674287
```

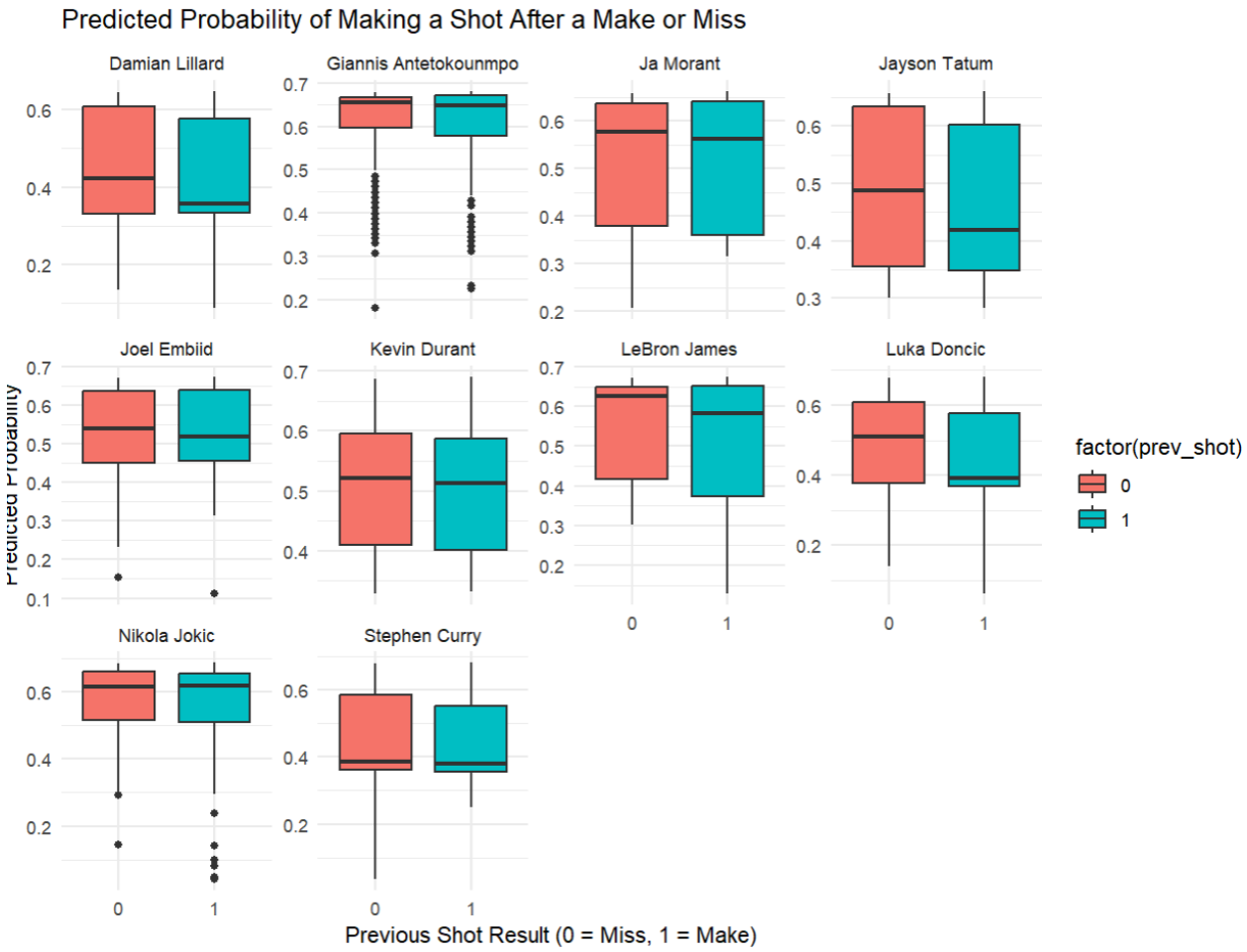
This shows the actual coefficient for the variable “prev_shot” and the result is a negative coefficient. This suggests that making the previous shot decreases the chances of making the next.

Abdullah Al Iman
City College of New York
Econ B2000
Prof. Kevin Foster
Dec. 2024



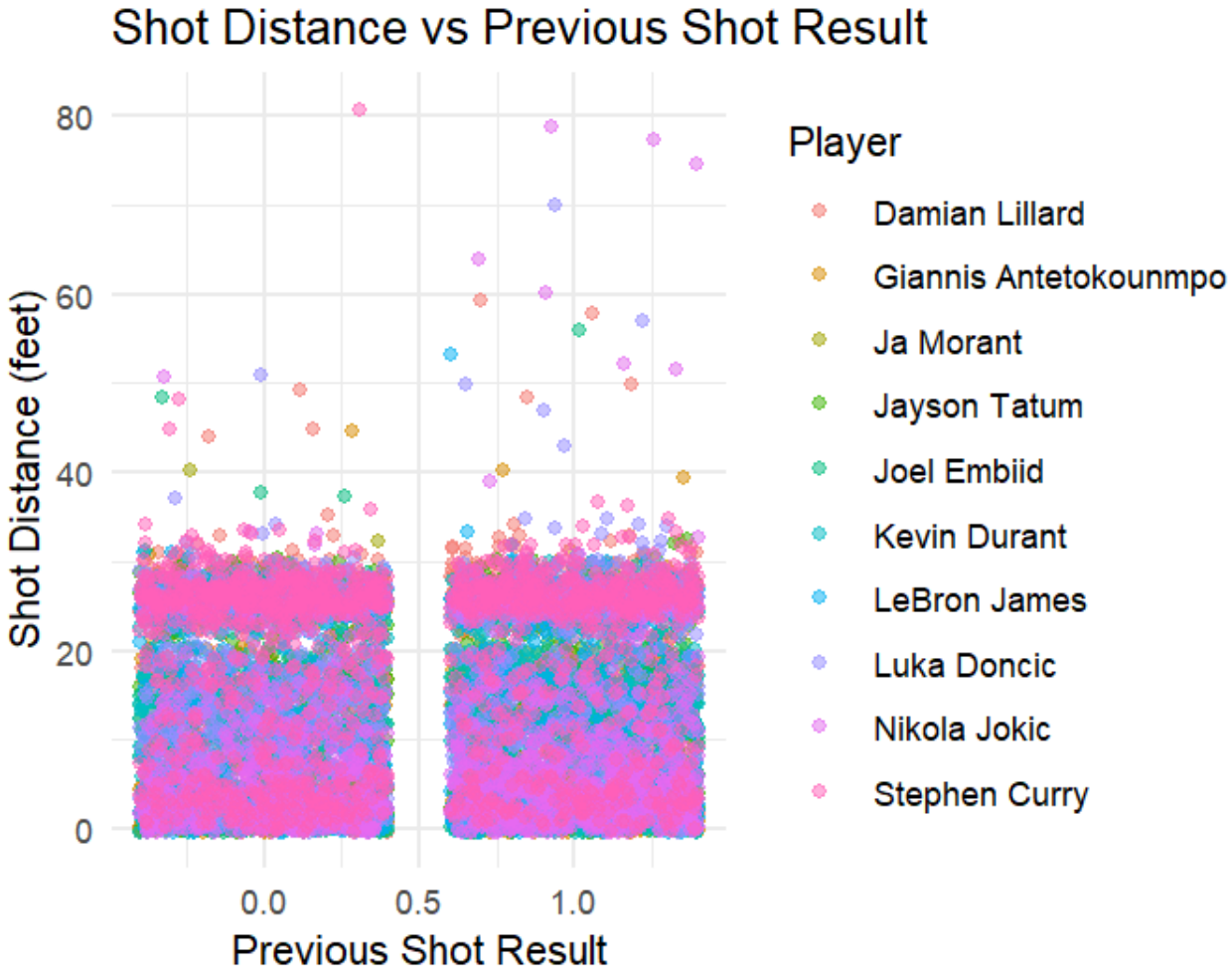
As we can see, the vertical red line falls to the far left of the histogram. The red line is the actual coefficient and the histogram is the distribution of coefficients by chance. Since the actual coefficient is much more negative than the data, we can conclude that making the previous shot decreases the chance of making the next. While that disproves the “Hot Hand” theory it actually rejects my null hypothesis as well because it seems as though the result of the previous shot has an impact on the next , just not in the way that was originally thought (decreases the chances instead of increasing). In other words, this data actually shows us that if the previous shot was made , there is a higher chance of getting a “Cold Hand”.

Abdullah Al Iman
City College of New York
Econ B2000
Prof. Kevin Foster
Dec. 2024



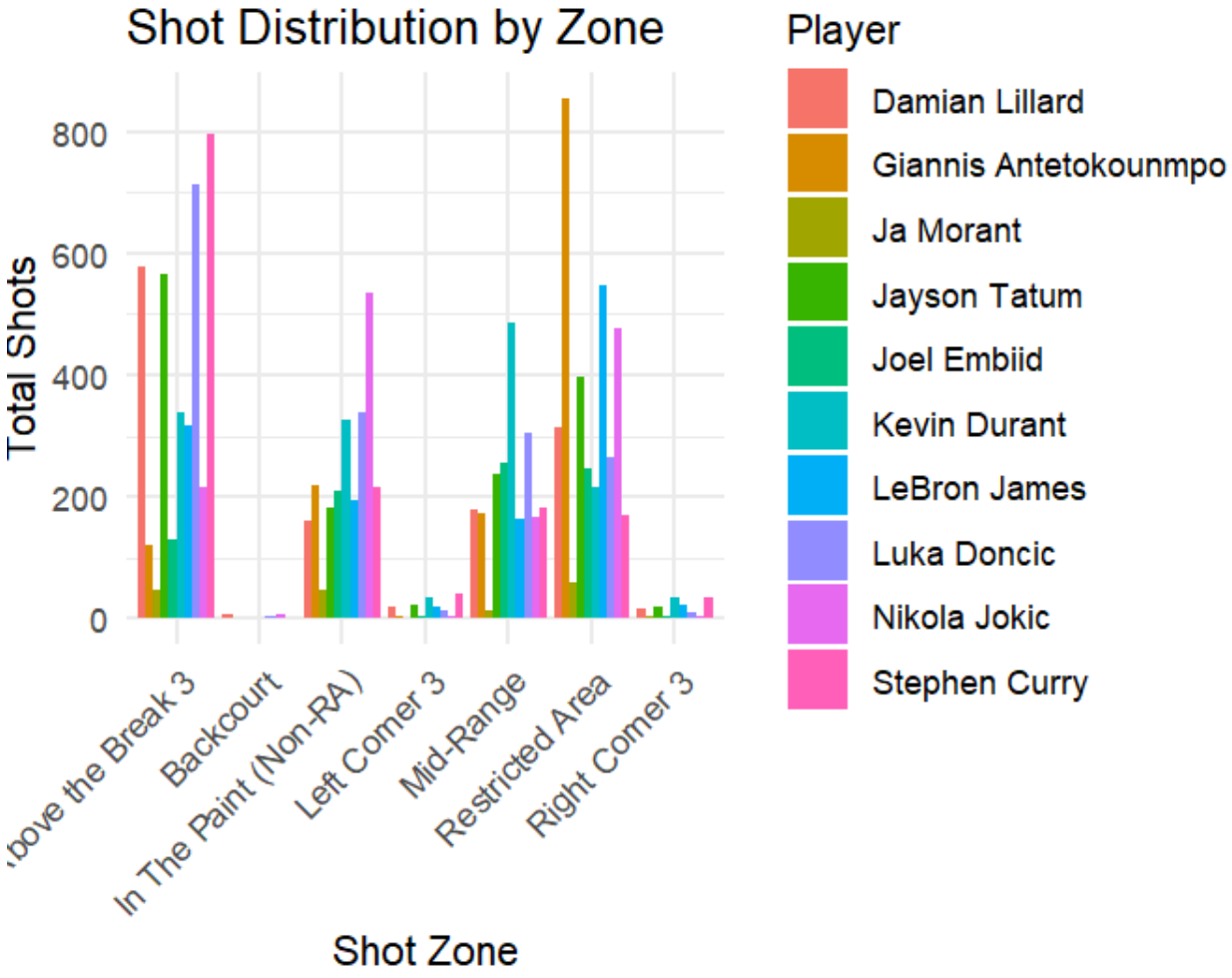
This graph gives a visual to the hot hand as it shows the probability of making a shot based on the previous make. The distributions for “0” and “1” overlap which shows no difference therefore further proving that the “hot hand” may not be true.

Abdullah Al Iman
City College of New York
Econ B2000
Prof. Kevin Foster
Dec. 2024



This graph shows the distance of a shot a player takes right after they made their previous shot. As it is shown , there is no major change in distances between a previous make or miss as the two graphs look identical.

Abdullah Al Iman
City College of New York
Econ B2000
Prof. Kevin Foster
Dec. 2024



This graph showcases the amount of shots from which area of the court each player takes. Some obvious notes as Giannis has the most attempts from the restricted area whereas Steph has the most from the 3 point line.

Overall, the data proved to reject my null hypothesis which stated that the results of the previous shot will have no impact on the next. It seems as though making the previous actually lessens the chances of a player making the next. Quite the opposite of the “Hot Hand”

Abdullah Al Iman
City College of New York
Econ B2000
Prof. Kevin Foster
Dec. 2024
phenomenon. This was evident by the logistic regression, monte carlo simulation and the visualizations.