Modeling Shot Efficiency in the NBA

Stat guys: Lewis Eatherton, Chris Yang, Charlie Bonetti

10/28/20

Your written report goes here! Before you submit, Make sure your code chunks are turned off with `echo = FALSE` and there are no warnings or messages with `warning = FALSE` and `message = FALSE`

Introduction

For our STA 210 Final Project, our group is interested in investigating how certain factors may have an influence on a basketball players' shot efficiency during a game in the National Basketball Association (NBA). We are all avid NBA enthusiasts and express great curiosity in what makes a good shot versus a bad shot. This is an exciting opportunity to explore the presence or absence of certain variables that surprisingly, or unsurprisingly, have an effect on scoring a basketball shot, also known as shot success.

Our interest in this subject matter was piqued by an 2015 article titled "Basketball Shot Types and Shot Success in Different Levels of Competitive Basketball" published by Frane Erčulj and Erik Štrumbel. The study described the effect of different shot types on shot success across different levels of competition. It was insightful to learn that there were no discernable differences between different situational variables, such as the type of player (e.g. Center) or where he shot the ball (e.g. in the paint) on shot success between levels. Since it was demonstrated that the effect of situation variables on shot success remained constant throughout all levels of competition, we were motivated to identify what exactly these situation variables might be, and to what degree of influence they had on an individual's shot success.

As a result, we formulated our research question: "Do certain situational variables during an NBA game have an effect on a basketball player's shot success?" Based on our prior knowledge of the game of basketball, we recognize that both players and coaches deem certain shots as "great" as "bad". Yet, what qualities characterize these shots to fall under those two distinct categories? Furthermore, through our own intuition after watching countless games and playing the sport ourselves, we hypothesize that some situational variables (e.g. shot distance from the hoop or the proximity of the closest defender) will have a greater effect on an NBA player's shot success than other situational variables. It seems intuitive that, as the time left on the clock and distance to the defender decrease, the likelihood of making a shot increases. To what extent is this true? Seeking clarity in how these factors, amongst others, contribute to a successful shot are the key objectives of our analysis.

Data

To seek answers to our series of questions, we will explore data from the 2014-2015 National Basketball Association (NBA) season. The data was originally collected via the NBA API PHP Library, a software tool developed by Jason Roman which scrapes historical information of past games, including statistical data, from previous seasons off of the NBA website. The data from the NBA website was acquired via manual documentation by analysts at the scoring table during games as well as computer vision techniques that track numbers such as scores, assists, etc.

Each observation in our data set is a unique shot that was taken in an NBA basketball game during the 2014-2015 season. Each observations is characterized by 21 features, including factors such as the game matchup, player name, and shot result. In total, there are 128,069 observations within the data set. For

our project, the response variable of interest is shot success. This value determines the whether or not a successful shot was made after all relevant predictor values were taken into consideration.

The response variable we are interested in investigating is:

- `SHOT_SUCCESS`: An indicator variable quantifying whether or not the shot was made.

The predictor variables that we expect to have an influence on the response variable are:

- `SHOT_NUMBER`: Shot number for a given player in a specific game (e.g. 8 denotes the player's 8th shot in game)
- `SHOT_CLOCK`: Number of seconds left on the shot clock when the player shot the ball
- `SHOT_DIST (FT)`: Distance in feet from the hoop when the player shot the ball
- `SHOT_RESULT`: Whether the shot was "made" or "missed"
- `DRIBBLES`: Number of dribbles before the player shot the ball
- `TOUCH_TIME (SEC)`: Number of seconds before the player shot the ball (after being passed to)
- `PTS_TYPE`: Whether the shot was 2 or 3-pointer
- `CLOSEST_DEFENDER`: Who the closest defender was to the shooting player
- `CLOSE_DEF_DIST`: Number of feet the defender was from the player when the ball was shot
- `Player_name`: Who shot the ball

```
## Rows: 128,069
## Columns: 21
## $ GAME_ID                     <dbl> 21400899, 21400899, 21400899, 21400899, ...
## $ MATCHUP                     <chr> "MAR 04, 2015 - CHA @ BKN", "MAR 04, 201...
## $ LOCATION                    <chr> "A", "A", "A", "A", "A", "A", "A", "A", ...
## $ W                           <chr> "W", "W", "W", "W", "W", "W", "W", "W", ...
## $ FINAL_MARGIN                <dbl> 24, 24, 24, 24, 24, 24, 24, 24, 24, 1, 1...
## $ SHOT_NUMBER                 <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 1, 2, 3, 4, 1...
## $ PERIOD                      <dbl> 1, 1, 1, 2, 2, 2, 4, 4, 4, 2, 2, 4, 4, 4...
## $ GAME_CLOCK                  <time> 01:09:00, 00:14:00, 00:00:00, 11:47:00,...
## $ SHOT_CLOCK                  <dbl> 10.8, 3.4, NA, 10.3, 10.9, 9.1, 14.5, 3....
## $ DRIBBLES                    <dbl> 2, 0, 3, 2, 2, 2, 11, 3, 0, 0, 8, 14, 2,...
## $ TOUCH_TIME                  <dbl> 1.9, 0.8, 2.7, 1.9, 2.7, 4.4, 9.0, 2.5, ...
## $ SHOT_DIST                   <dbl> 7.7, 28.2, 10.1, 17.2, 3.7, 18.4, 20.7, ...
## $ PTS_TYPE                    <dbl> 2, 3, 2, 2, 2, 2, 2, 2, 3, 3, 3, 2, 2, 3...
## $ SHOT_RESULT                 <chr> "made", "missed", "missed", "missed", "m...
## $ CLOSEST_DEFENDER            <chr> "Anderson, Alan", "Bogdanovic, Bojan", "...
## $ CLOSEST_DEFENDER_PLAYER_ID <dbl> 101187, 202711, 202711, 203900, 201152, ...
## $ CLOSE_DEF_DIST              <dbl> 1.3, 6.1, 0.9, 3.4, 1.1, 2.6, 6.1, 2.1, ...
## $ FGM                         <dbl> 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 0...
## $ PTS                         <dbl> 2, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 2, 2, 0...
## $ player_name                 <chr> "brian roberts", "brian roberts", "brian...
## $ player_id                   <dbl> 203148, 203148, 203148, 203148, 203148, ...
```

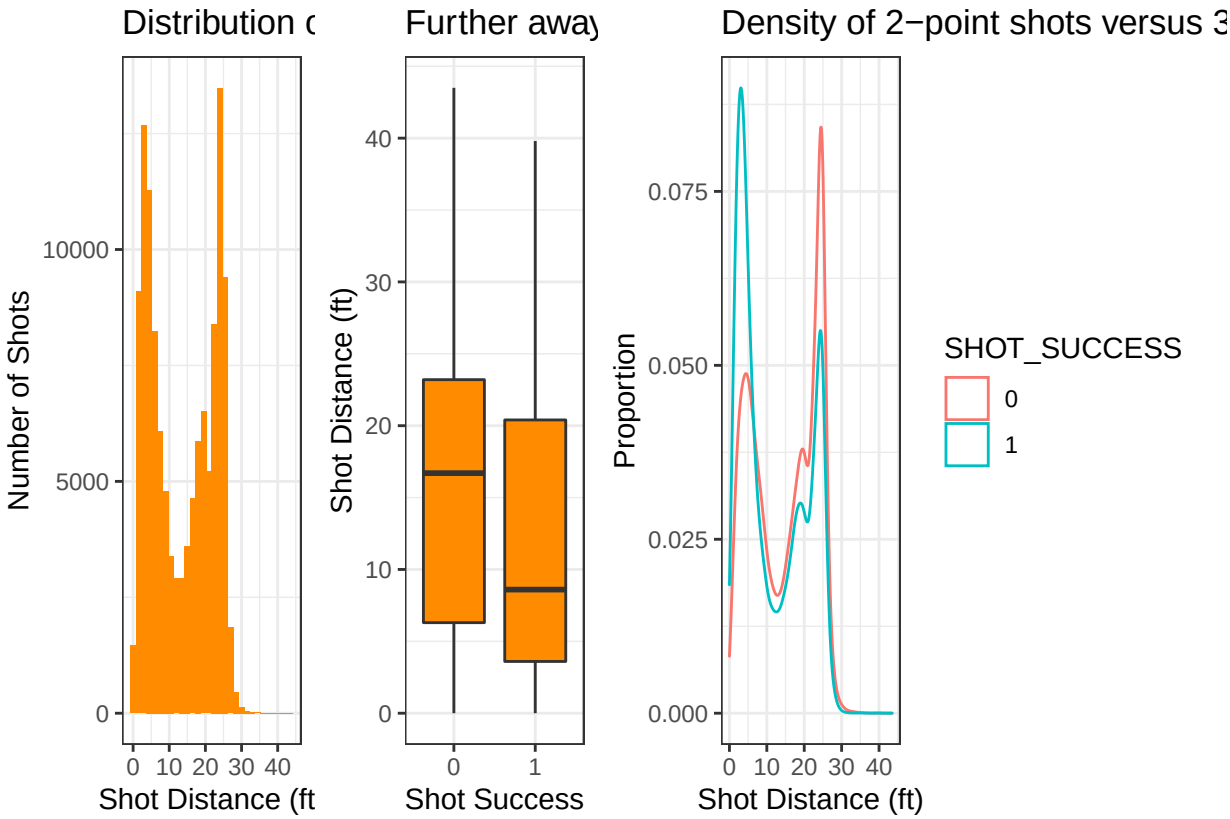
Data Cleansing

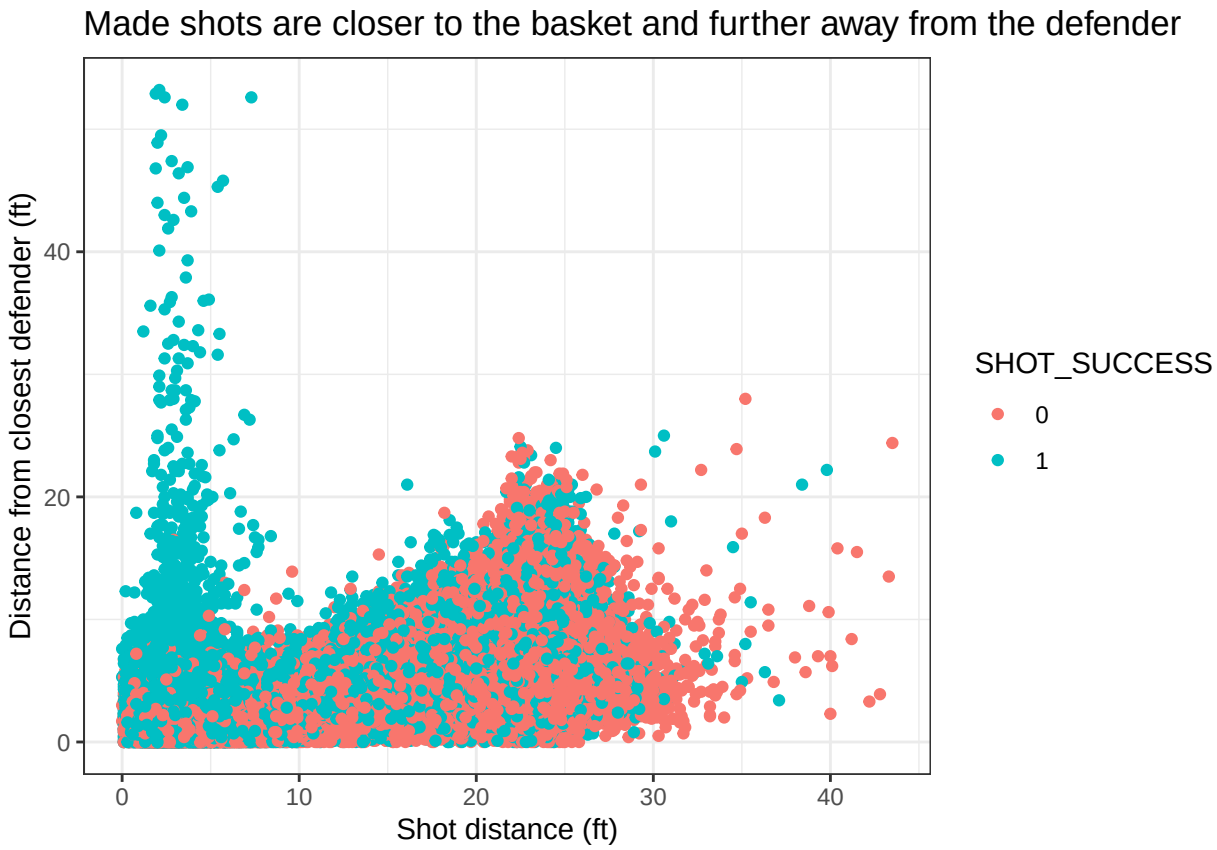
```
##                    GAME_ID                     MATCHUP
##                          0                           0
##                   LOCATION                           W
##                          0                           0
##               FINAL_MARGIN                 SHOT_NUMBER
##                          0                           0
##                     PERIOD                 GAME_CLOCK
##                          0                           0
##                 SHOT_CLOCK                    DRIBBLES
##                       5567                           0
```

```
##                  TOUCH_TIME                   SHOT_DIST
##                           0                           0
##                   PTS_TYPE                 SHOT_RESULT
##                           0                           0
##           CLOSEST_DEFENDER CLOSEST_DEFENDER_PLAYER_ID
##                           0                           0
##              CLOSE_DEF_DIST                         FGM
##                           0                           0
##                         PTS                 player_name
##                           0                           0
##                   player_id                SHOT_SUCCESS
##                           0                           0
```

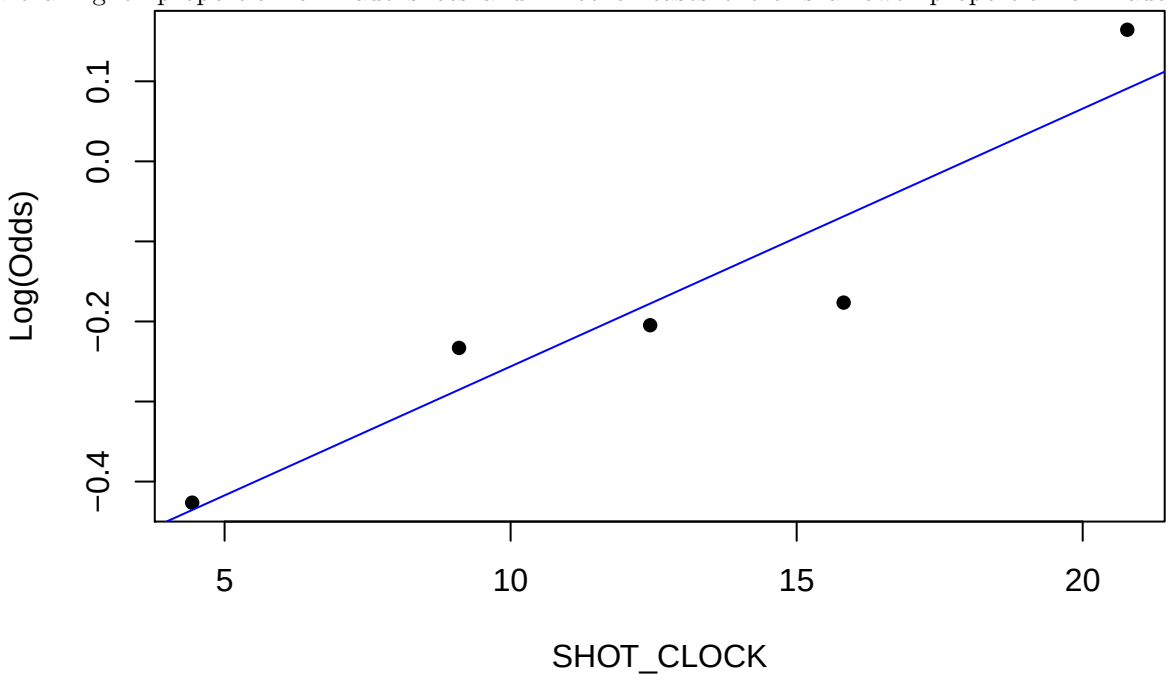
Analysis of missing data reveals that only one variable contains unavailable values. To accommodate for this, we removed 5567 missing values from the dataset since they accounted for only 4.3468755 percent of all observations, which is an inconsequential amount.

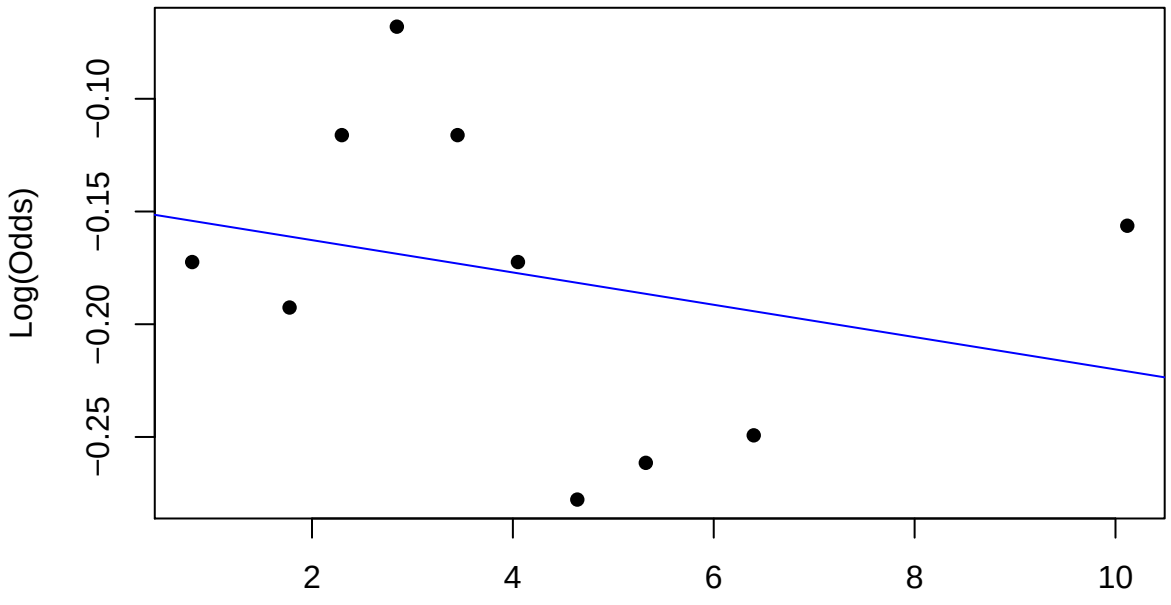
Exploratory Data Analysis

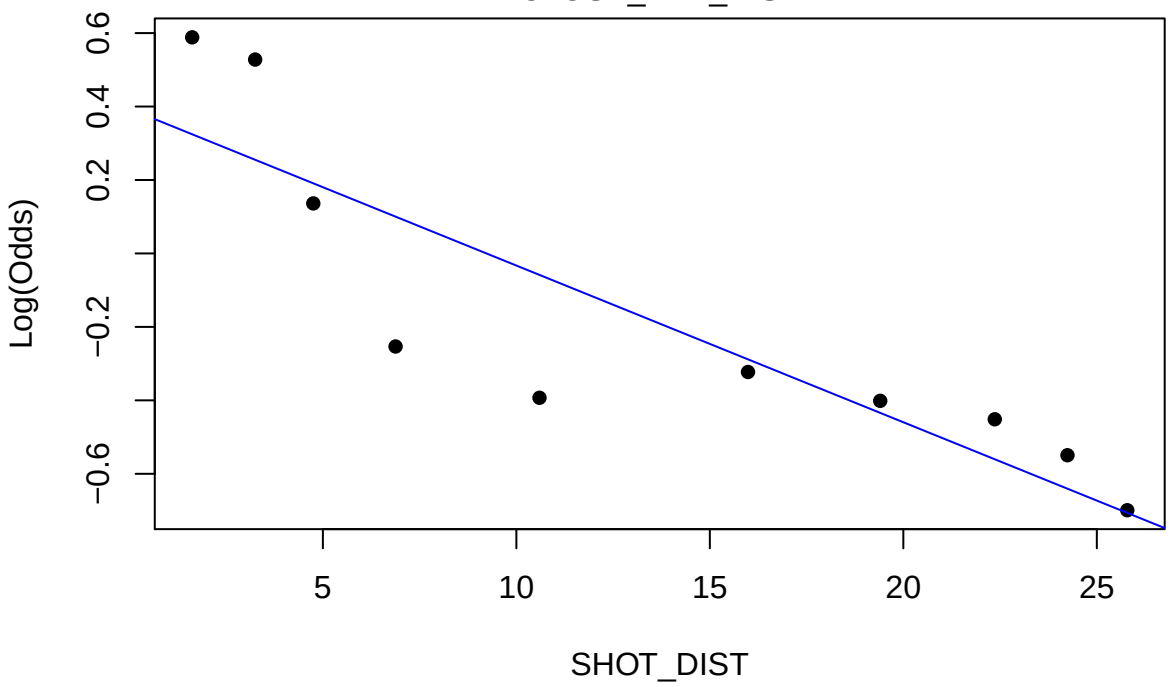


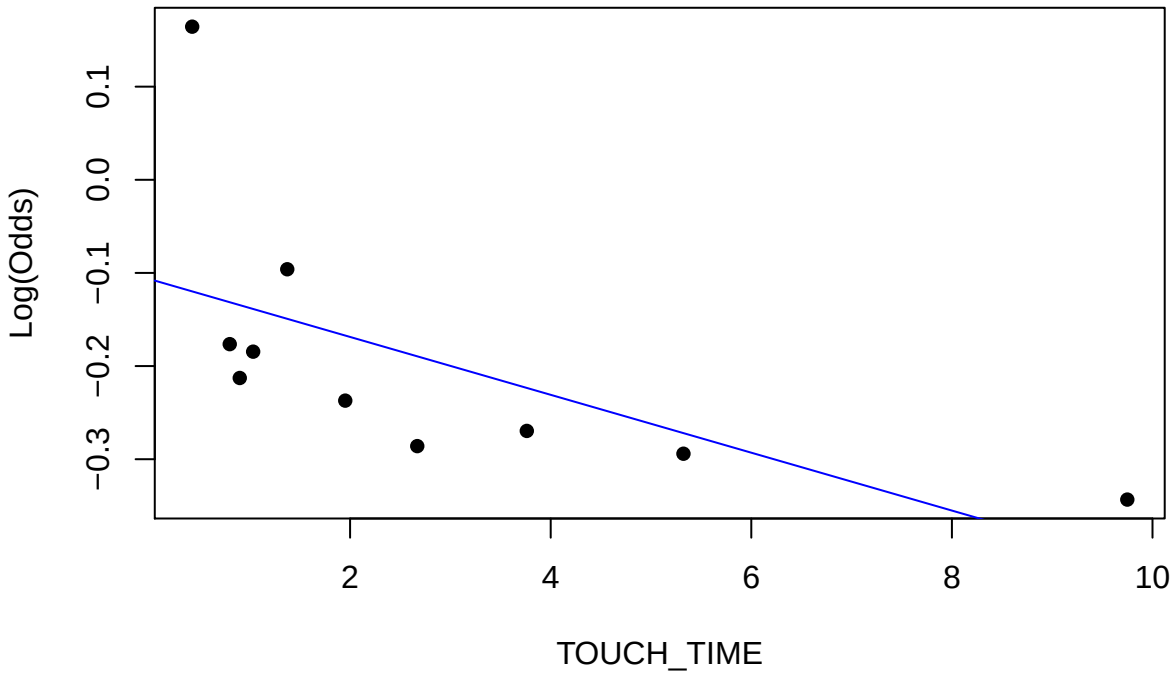


According to this plot, there appears to be a relationship between shot distance and shot success, but not distance from defender and shot success. In the plot, there are clearly more made shots than not made shots when the shot is close to the basket, but as the shot distance goes further away there is a significant increase in proportion of missed shots as compared to made shots. On the other hand, the distance from closest defender does not appear to tell us much about the shot success as in some cases there are a higher proportion of made shots and in other cases there is a lower proportion of made shots.

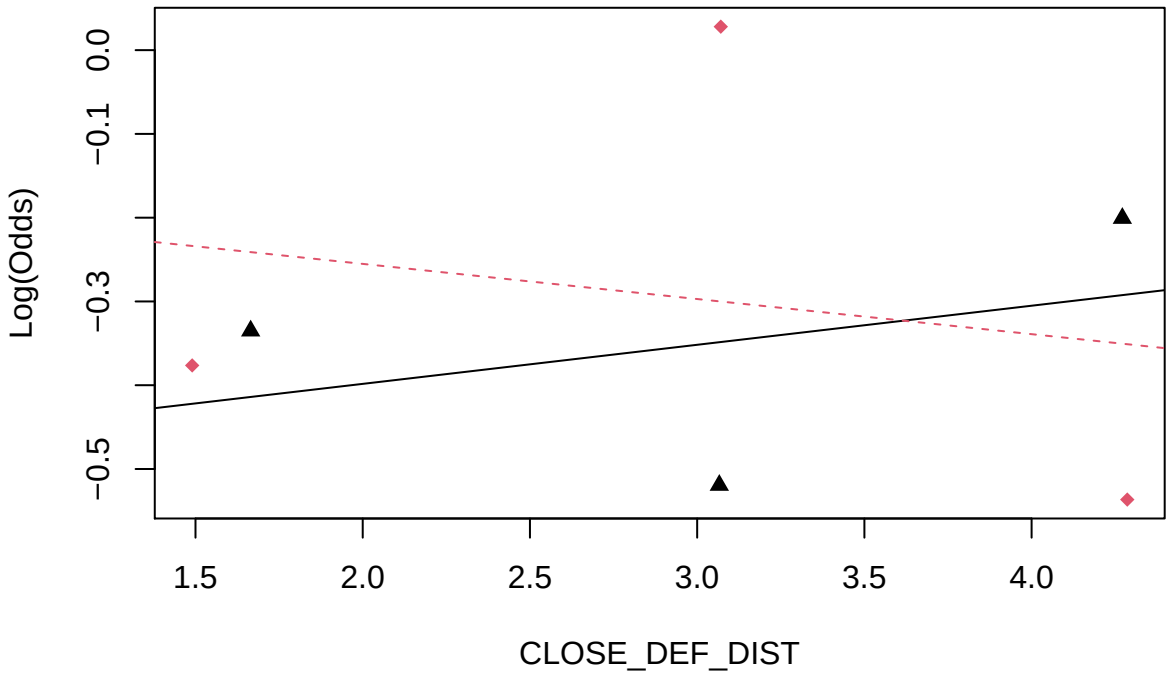



Log(Odds)
−0.10
−0.15
−0.20
−0.25
2
4
6
8
10
CLOSE_DEF_DIST


Log(Odds)
0.6
0.4
0.2
−0.2
−0.6
5
10
15
20
25
SHOT_DIST




```
## # A tibble: 469 x 5
## # Groups:   CLOSEST_DEFENDER [469]
##    CLOSEST_DEFENDER   SHOT_SUCCESS     n  prop emp_logit
##    <chr>              <fct>        <int> <dbl>     <dbl>
##  1 Acy, Quincy        1              117 0.429    -0.288
##  2 Adams, Jordan      1               15 0.536     0.143
##  3 Adams, Steven      1              212 0.443    -0.231
##  4 Adrien, Jeff       1               39 0.549     0.198
##  5 Afflalo, Arron     1              181 0.420    -0.323
##  6 Ajinca, Alexis     1              111 0.474    -0.103
##  7 Aldemir, Furkan    1               33 0.465    -0.141
##  8 Aldrich, Cole      1              141 0.532     0.128
##  9 Aldridge, LaMarcus 1              296 0.469    -0.124
## 10 Allen, Lavoy       1              134 0.447    -0.214
## # ... with 459 more rows
```



Linearity condition is satisfied for the variables of SHOT_CLOCK and SHOT_DIST because observations appear to be symettric across the blue line. However, linearity condition for the variables CLOSE_DEF_DIST and TOUCH_TIME does not appear to be satisfied because observations do not appear to be symmetric across the line.

The independence condition is satisfied because the shot of one person does not appear to have an affect on the shot of another person.

The condition for randomness is satisfied becuase we have no reason to beleive otherwise when considering how the data was collected.

we should also look into collinearity

Creating Model

```
## Single term deletions
##
## Model:
## SHOT_SUCCESS ~ SHOT_CLOCK + DRIBBLES + TOUCH_TIME + SHOT_DIST +
##     CLOSE_DEF_DIST
##                Df Deviance    AIC
## <none>              162325 162337
## SHOT_CLOCK      1   162532 162542
## DRIBBLES        1   162354 162364
## TOUCH_TIME      1   162434 162444
## SHOT_DIST       1   167561 167571
## CLOSE_DEF_DIST  1   163764 163774
```

```
## Single term deletions
##
## Model:
## SHOT_SUCCESS ~ SHOT_CLOCK + DRIBBLES + TOUCH_TIME + SHOT_DIST +
##     CLOSE_DEF_DIST + CLOSEST_DEFENDER
##                   Df Deviance    AIC
## <none>                 855.93 869.93
```

```
## SHOT_CLOCK        1   858.72 870.72
## DRIBBLES          1   855.93 867.93
## TOUCH_TIME        1   855.94 867.94
## SHOT_DIST         1   872.34 884.34
## CLOSE_DEF_DIST    1   865.25 877.25
## CLOSEST_DEFENDER  1   855.98 867.98
```

```
## Single term deletions
##
## Model:
## SHOT_SUCCESS ~ SHOT_CLOCK + TOUCH_TIME + SHOT_DIST + CLOSE_DEF_DIST +
##     CLOSEST_DEFENDER
##                   Df Deviance    AIC
## <none>                 855.93 867.93
## SHOT_CLOCK         1   858.81 868.81
## TOUCH_TIME         1   855.99 865.99
## SHOT_DIST          1   872.34 882.34
## CLOSE_DEF_DIST     1   865.25 875.25
## CLOSEST_DEFENDER   1   855.98 865.98
```

```
## Single term deletions
##
## Model:
## SHOT_SUCCESS ~ SHOT_CLOCK + TOUCH_TIME + SHOT_DIST + CLOSE_DEF_DIST
##                 Df Deviance    AIC
## <none>               855.98 865.98
## SHOT_CLOCK       1   858.90 866.90
## TOUCH_TIME       1   856.02 864.02
## SHOT_DIST        1   872.45 880.45
## CLOSE_DEF_DIST   1   865.36 873.36
```

```
## Single term deletions
##
## Model:
## SHOT_SUCCESS ~ SHOT_CLOCK + SHOT_DIST + CLOSE_DEF_DIST
##                 Df Deviance    AIC
## <none>               856.02 864.02
## SHOT_CLOCK       1   859.04 865.04
## SHOT_DIST        1   872.60 878.60
## CLOSE_DEF_DIST   1   865.68 871.68
```

```
## # A tibble: 2 x 5
##   Resid..Df Resid..Dev    df Deviance p.value
##       <dbl>      <dbl> <dbl>    <dbl>   <dbl>
## 1       645       856.    NA   NA      NA
## 2       643       855.     2    0.813   0.666
```

term	estimate	std.error	statistic	p.value
(Intercept)	-0.737	0.294	-2.508	0.012
SHOT_CLOCK	0.026	0.015	1.733	0.083
SHOT_DIST	-0.048	0.012	-4.013	0.000
CLOSE_DEF_DIST	0.243	0.079	3.067	0.002

talk about final model outcome and how we came to it ***(Not sure what goal of final model is - ex. prediction or explanation or one variable's effect)

In creating our final model we chose to conduct a backward selection procedure. Therefore, we started out with the model that included all variables that we assumed to have a significnat effect on the response variable, in this case shot success. The predictor variables we selected in our initial model were SHOT_CLOCK, CLOSEST_DEFENDER, TOUCH_TIME, SHOT_DIST, CLOSE_DEF_DIST. In terms of conducting the backward selection, we eliminated the variables based on which predictor variable resulted in the model having the lowest AIC value. When we could not eliminate anymore predictor variables from the model based on this criteria, we thus arrived on our final model. The predictors in our final model were SHOT_CLOCK, SHOT_DIST, and CLOSE_DEF_DIST. Our final model can be expressed as:

predicted SHOT_SUCCESS ~ -.737 + .026 * SHOT_CLOCK - .048 * SHOT_DISTANCE + .243 * CLOSE_DEF_DIST

###Discussion

Limitations:

1) Out of all predictor variables in analysis, only one variable had missing observations; there were 5,567 unavailable values for the variable which counted the time left on the shot clock when the player shot the ball. This may have been due to recording error or instances where the shot clock was turned off. Given that 5,567 is marginal compared to the total number of observations (128,069), we chose to remove all rows containing NAs for this variable from the dataset. As a result, The dataset wouldn't differ to a significant degree systematically compared to the larger population.
2) The variable which quantified the number of seconds where a player touched the ball was negative

Next steps:

1) To address the limitation of _____,
2) Furthermore, an avenue we would be interested in exploring how the situational variables of making a successful shot may differ or stay the same during "clutch" scenarios, where the outcome of a game is at stake. In this case, we could explore other predictors within our model and set thresholds, such as having less than 10 seconds left during the game, being in the 4th period of the game, and having a final margin of victory of less than or equal to 3 points. Furthermore, another avenue we would be interested in exploring is the effectiveness of different defenders on a player's shot success in NBA games. In the current stage of our project, we investigated the variables that played in a role in optimizing a successful shot. Out of curiosity, how would these factors change if we explored the data from the perspective of the defender, who is attempting to prevent a player from scoring a basket? Further analysis of the distance between contested shots as well as field goal percentage could yield interesting factors that may have some degree of influence on defending a shot. Additionally, based on these characteristics and others, perhaps we could identify which NBA players constitute the "elite" class of the defensive side of basketball.

Talk about our results, the limitations of these results, and what'd we do differently...

###References

(1) Erčulj F, Štrumbelj E (2015) Basketball Shot Types and Shot Success in Different Levels of Competitive Basketball. PLoS ONE 10(6): e0128885. https://doi.org/10.1371/journal.pone.0128885