Baseball and Data Analysis: Height and Hitting Power

Chris Kucewicz

Does height help predict hitting power?

Baseball was my passion as a child. I was an avid fan and played competitively on several travel teams. One thing that limited my playing career though, was my height. It was difficult to get noticed by big-time college coaches because of how short I am. An unwritten rule of thumb is that in order to get a Division 1 scholarship or get drafted you had to be at least 6 feet tall or have olympic level speed.

This was discouraging because I knew I could compete with the same players that were taller, so this led me to ask the question: “Does height help predict performance in baseball?” One of the measures of performance in baseball is hitting with power, and the hit that demonstrates the most power is a home run. So in this analysis project, I compared the heights of all major league baseball players against their career home run total to see how much height predicts hitting performance. Specifically how well does height predict a player’s power?

Steps

1. Gather data and install necessary packages
2. Compile the career home runs for each player
3. Plot each player's height against their career home runs
4. Find the value of correlation coefficient
5. Analyze relationship between player height and home run totals

Resources:
"The R Series: Analyzing Baseball Data with R" by Max Marchi and Jim Albert

1. Gather data and install necessary packages

For this project, I used datasets from the Lahman database, which is a database that contains “complete batting and pitching statistics from 1871 to 2020, plus fielding statistics, standings, team stats, managerial records, post-season data, and more.” Once downloaded, the two datasets I loaded were the ‘Batting’ dataset, which I used to compile the career home runs for each player, and the ‘People’ dataset, which contained the heights for all players.

Since I was working across multiple datasets, I installed the ‘plyr’ library. The ‘ddply’ function in this library is useful for splitting and combining datasets.

Code:

```
setwd("C:/Users/ckuce/Data Analytics/Data
Analytics Scripts/R/win-library/3.5")

Batting <- read.csv("C:/Users/ckuce/Data
Analytics/Data
AnalyticsScripts/R/win-library/3.5/baseballdata
bank-master_2018-03-28/baseballdatabank-master/
core/Batting.csv")

People <- read.csv("C:/Users/ckuce/Data
Analytics/Data
AnalyticsScripts/R/win-library/3.5/baseballdata
bank-master_2018-03-28/baseballdatabank-master/
core/People.csv")

install.packages('plyr')

library(plyr)
```

2. Compile career home runs for each player

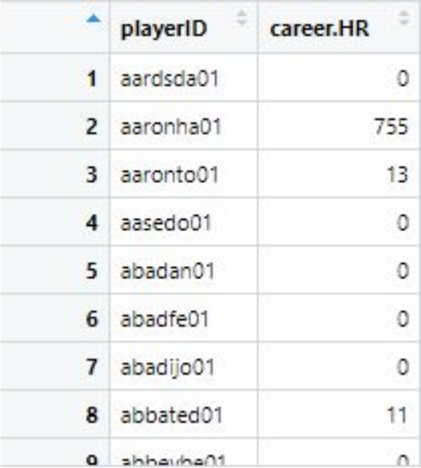
The ‘Batting’ dataset gave the yearly home run total for each player. To get each player’s career home run total, I used the ‘ddply’ function to add the yearly home runs for each player each year that they played.

With the code below, R created a new dataframe with the career home runs for each player, so I used the ‘merge’ function to combine the ‘Batting’ dataframe with this dataframe as a new variable

Code:

```
dataframe.HR <- ddply(Batting, .(playerID),
summarize, career.HR = sum(HR, na.rm = TRUE))


Batting <- merge(Batting, dataframe.HR, by =
"playerID")
```

dataframe.HR



	playerID	career.HR
1	aardsda01	0
2	aaronha01	755
3	aaronto01	13
4	aasedo01	0
5	abadan01	0
6	abadfe01	0
7	abadijo01	0
8	abbated01	11
9	abbeybe01	0

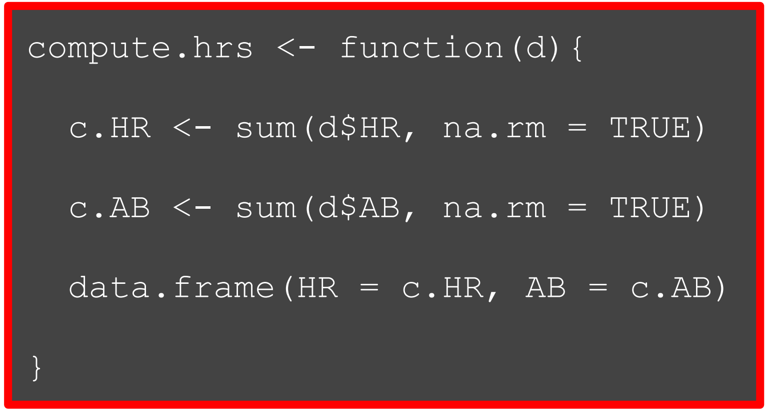
The above image shows the first 8 rows of the dataframe.HR data frame

2. Compile career home runs for each player

One of the dilemmas up to this point was that my code had been analyzing the home run totals for every single player even those who had as little as one career at bat in Major League Baseball. This made the run time for my code longer as it had to process thousands of more entries.

In order to make my code more efficient and my analysis more reliable, I decided to only analyze the players who had 5000 or more career at bats, which is roughly 9 full seasons worth of at bats. To compile a dataframe of the career at bats and career home runs for each player, I created a function called 'compute.hrs' (shown on the right) which computes the total at bats and total home runs for each player. The entire code is shown on the following slide.

compute.hrs
function:



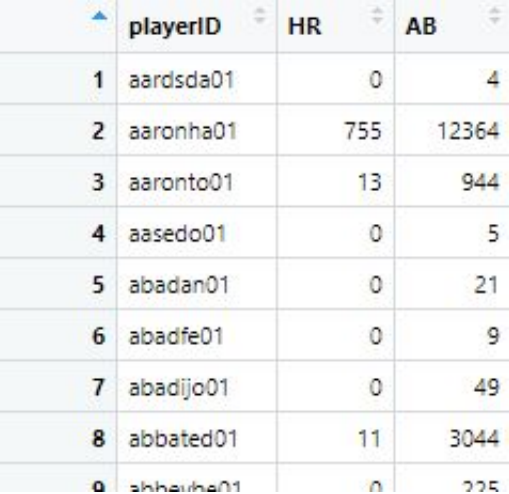
```
compute.hrs <- function(d){

  c.HR <- sum(d$HR, na.rm = TRUE)

  c.AB <- sum(d$AB, na.rm = TRUE)

  data.frame(HR = c.HR, AB = c.AB)

}
```

2. Compile career home runs for each player

Code:

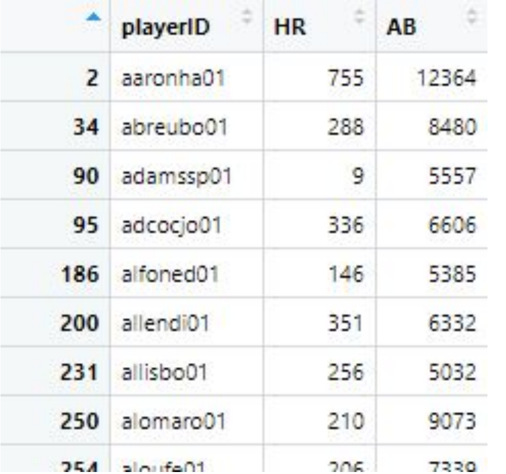
```
source("compute.hrs.R")

d.HR.AB <- ddply(Batting, .(playerID),
compute.hrs)

AB.5000 <- subset(d.HR.AB, AB >= 5000)

player.heights <-
data.frame(People$playerID,
People$height)

colnames(player.heights)[1] <- "playerID"

Height.HR.5000ab <- merge(AB.5000,
player.heights, by = "playerID")

Height.HR.5000ab$AB <- NULL
```

d.HR.AB dataframe



	playerID	HR	AB
1	aardsda01	0	4
2	aaronha01	755	12364
3	aaronto01	13	944
4	aasedo01	0	5
5	abadan01	0	21
6	abadfe01	0	9
7	abadijo01	0	49
8	abbated01	11	3044
9	abbeybe01	0	225

AB.5000 dataframe



	playerID	HR	AB
2	aaronha01	755	12364
34	abreubo01	288	8480
90	adamssp01	9	5557
95	adcocjo01	336	6606
186	alfoned01	146	5385
200	allendi01	351	6332
231	allisbo01	256	5032
250	alomaro01	210	9073
254	aloufe01	206	7339

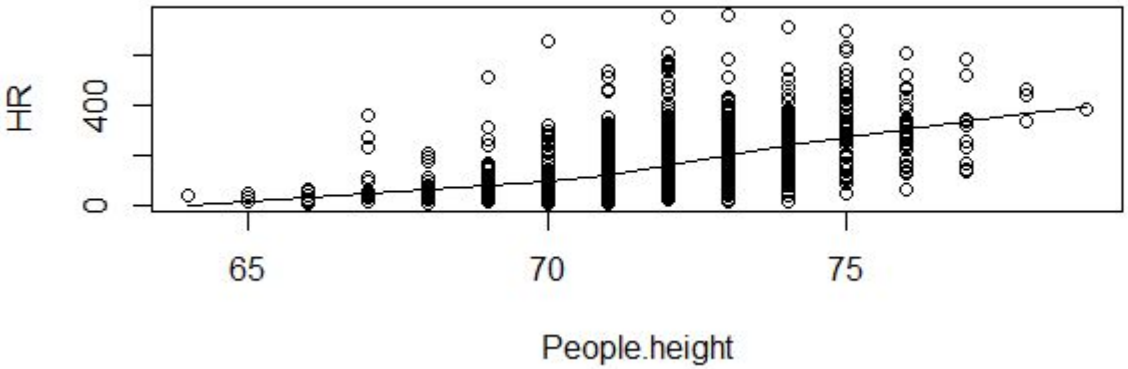
The d.HR.AB dataframe shows the home runs (HR) and at bats (AB) for all players, while the AB.5000 data frame includes only the players with at least 5000 career at bats.

3-4. Plot each player’s height against their career home runs, and find the value of correlation coefficient

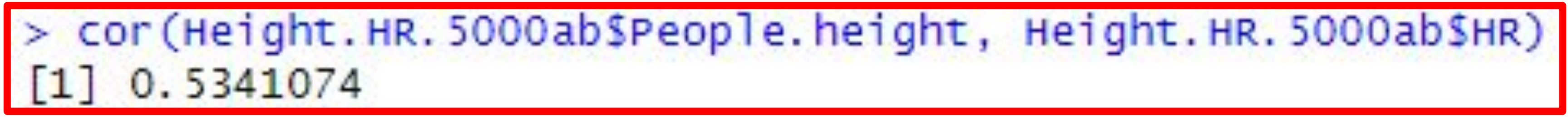
Code:

```
with(Height.HR.5000ab,
plot(People.height, HR))

with(Height.HR.5000ab,
lines(lowess(People.height,
HR)))

cor(Height.HR.5000ab$People.hei
ght, Height.HR.5000ab$HR)
```

Graph of Players’ height (inches) vs. Career Home Runs



Output:



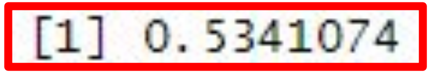
```
> cor(Height.HR.5000ab$People.height, Height.HR.5000ab$HR)
[1] 0.5341074
```

The graph above shows a positive relationship between player height and number of home runs. The correlation coefficient has a value of 0.534 which indicates that there is a weak, positive correlation between player height and the number of career home runs.

5. Analysis and implications

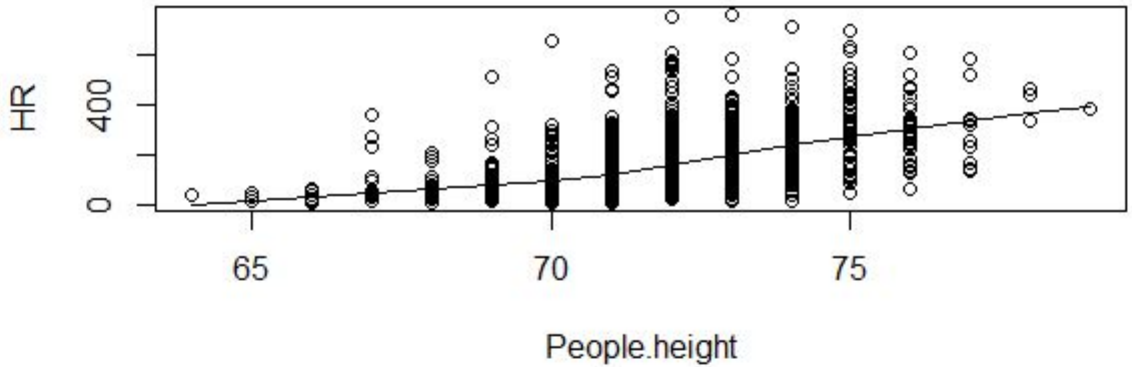
Although the graph shows a slightly positive relationship between player height and their career home runs, the correlation coefficient value of 0.534 indicates that there is not enough of a correlation between the height of a player and their hitting power to be a statistically significant relationship.

Value of the correlation coefficient



```
[1] 0.5341074
```

Graph of Players’ height (inches) vs. Career Home Runs



The implication for this analysis means that when scouts and coaches are deciding on which players to make up their team, they should not solely use height as a measure to evaluate the player’s ability to hit the ball hard and far. There are other factors which play a role in a player’s hitting power, such as their bat speed, overall strength, and how often their swing makes solid contact with the pitch.

Reflection

- **What improvements or changes would you make if you had to do this again?**
  - One change I would make would be to improve the scatter plot's readability by including clearer axis labels and scaling the y-axis to improve the visibility of the data points.
  - I might also use a different statistic to measure a player's hitting power. Instead of home runs, I might use slugging percentage or the amount of extra base hits. Hitting doubles and triples are ways to demonstrate hitting power that home runs does not account for. Both slugging percentage and the number of extra base hits take doubles and triples into account.
- **What did you learn through your work on this project?**
  - Through this project I learned a great deal of R's syntax and functionality such as how to import datasets, how to split and combine existing data frames using the 'ddply' function in the 'plyr' package, how to create and run my own functions, how to create a scatter plot, how to find descriptive statistics of data specifically the correlation coefficient.
  - In addition to learning syntax, I learned the process of putting together a data analytics report.
- **Where does this project lead you?**
  - I was excited to work on my first data analytics project (and my first project in R) using a question that held personal significance to me being that I was overlooked by a lot of coaches because of my height during my playing days. I'm looking forward to continuing to grow in my knowledge of R syntax in order to create clearer visualizations and machine learning models that can help me to predict values such as the season win-loss records for teams.