WILLIAM & MARY

CHARTERED 1693

Swinging For the Fences: Analysis of Major League Baseball Home Runs

May 2024

Executive Summary

- In this analysis I will be looking at a variety of metrics to see what causes more home runs in Major League Baseball (MLB), such as exit velocity, launch angle, stadium friendliness, and weather.
- The home run dataset I used comes from the website data world (https://data.world). I later merged a weather dataset which I got from the baseballr package available in R.
- I am assuming that both of these data sets are providing accurate, unbiased information that will allow me to draw appropriate conclusions.

Notable Findings:

- The majority of home runs are fly balls (43,908) occurring between 22.5-32.5 degrees of launch angle
- The bulk of home runs occur at exit velocities from 100-110 mph.
- The average home run in the dataset only goes out in 23/30 MLB parks - The top 2 most hitter friendly parks in terms of total home runs hit are Oriole Park at Camden Yards and Great American Ballpark (Reds). The most unfriendly is PNC Park (Pirates).
- Temperature doesn't have a consistent effect on home run hitting
- Type of day seems to matter for home run hitting (sunny vs. cloudy, etc.)

Limitations:

- It would have been helpful to have an overall at bat dataset alongside my only home run dataset
- It was difficult to get a definitive answer to the most and least friendly ballparks because a lot of new stadiums were being built in the date range of our data.

Overview

My final project delves into MLB statistics, particularly home runs, aiming to uncover valuable insights in baseball. With the rise of analytics, every major league team now has its own analytics department, highlighting the profound impact data-driven strategies have on team performance and financial efficiency. Uncovering answers to my research questions on home runs could thus provide a potential competitive edge.

My research questions are:

- What is the optimal exit velocity and launch angle for hitting home runs?
- What stadiums are the most hitter friendly/have the most home runs - does weather have an impact?

Data Source and Assumptions

I got my dataset from Data World (https://data.world), a platform where users upload datasets used for projects. Similar to Kaggle, Data World fosters a community feel and offers valuable insights, including data types and descriptive statistics. Given all the above, I trust Data World as a reliable source. I got my weather dataset from the baseballr package in R. Being that it is a built in package to the software, and it matches my other data, I have good reason to trust it. I am assuming all data is accurate and unbiased.

My home run dataset includes every home run in the MLB from 2006-2017, for a total of 60,576 observations. There are 19 variables such as game date, ballpark, distance, exit velocity, launch angle, and more. I got my weather data set from a baseballr package, available by the following commands:

- library(baseballr)
- baseballr::get_game_info_sup_petti()

It has a record of every MLB game from 2008-2022, with 18 variables including game_date, ballpark (venue_name), temperature, and day type (other_weather). Since it had overlapping game date and ballpark variables with my home run dataset, I was able to merge them and get a sense for how weather may impact home runs.

Data Cleaning Process

To clean the home run dataset I dropped unneeded columns and renamed the remaining. I created a variable to assign different contact types to each home run based on its launch angle. I edited problematic/missing data for ease of use and created a stadium friendliness variable. I then prepared my weather dataset to merge with my home run dataset by changing the name, data types, and capitalization of like columns so that they matched and were ready for merging. After merging, I selected relevant columns I would be using for the analysis, and omitted NA weather values.

Notable Findings

Finding 1. As seen in Table 1, the majority of home runs are fly balls. On average, line drive home runs have a launch angle of ~23, with fly ball home runs at ~30. Figure 1 provides a visual of the launch angle distribution - most home runs occur between a launch angle of 22.5-32.5.

Finding 2. Figure 2 is a scatter plot of exit velocity vs. distance of home runs. We observe a strong positive relationship, meaning exit velocity is an important indicator of distance (and thus home run) potential. The red dashed lines represent the 90th percentile of each variable. Table 2 shows how higher exit velocity intervals are associated with longer home run distances, supporting our findings in Figure 2.

Finding 3. Figure 3 shows that the average home run in this dataset only goes out in 23/30 MLB parks from its red dashed line. This leads us to Table 3, where we can see the the most hitter friendly parks are Oriole Park at Camden Yards and Great American Ballpark (Reds). We can see that the unfriendly ones are new stadiums, so this isn't a good representation of their true friendliness. Getting rid of these new stadiums, the least friendly stadium that was around for the entire duration of our analysis is PNC Park (Pirates).

Finding 4. Figure 4 shows the proportional differences of temperature in all games vs. the temperature on days home runs were hit. If the home run bar is higher than the all games bar, then home runs are hit in a higher proportion on days of that temperature. For example, at the highest point of 72F, we can look at the direct comparison in Table 4, where 7.1% of home runs were hit at this temperature, compared to only 6.9% of games being played at this temperature. This means that home runs have a higher tendency of being hit at 72F. However, there doesn't seem to be a particular pattern of more home runs being hit during certain temperatures as seen in the difference column, which makes it seem like these results are fairly random.

Finding 5. Figure 5 tells us the type of day seems to matter when it comes to home run occurrence. It compares the percentage of home runs hit on certain day types compared to the overall games played on certain day types. Table 5 depicts this numerically, where partly cloudy, cloudy, and roof closed days each have a 1.5-2.1 higher home run occurrence percentage than games played percentage on that day type. On the other hand, a sunny day seems to produce a massive penalty to home run hitting as evidenced by the differential.

Limitations

Having a dataset of every at bat during 2006-2017 alongside my home run dataset from that time period would've improved my findings. Comparing home runs and their associated metrics with other types of at bat results for further robustness checks would've been useful. This is something to consider for the future.

Determining the most and least friendly stadiums proved difficult because a lot of new stadiums were being built in the date range of my data. Thus, there were a handful of stadiums with relatively less years compared to others in the dataset, so their friendliness score wasn't accurate. If I could have found a time period with very little stadium turnover rates, the results for this would've been clearer and more accurate.

Method

I am going to create several graphs (histograms, scatter plots, and bar plots) as well as tables to aid in answering my research questions because it is very helpful to visualize this type of data. I will be performing statistical analyses along the way to improve the interpretations of these visuals. I will also conduct hypothesis tests for claims on how weather effects home run hitting.

Data Analysis

```
# Read in data...  2006-2017 MLB home run data
hr.df.read <- read.csv("https://query.data.world/s/ocmu3u2i5thmhuevcrvxhcudoxe3ms?dws=00000"
    header = TRUE, stringsAsFactors = FALSE)

# View structure
str(hr.df.read)
```

```
'data.frame':   60576 obs. of  19 variables:
 $ RECORDID         : int  60576 60575 60574 60573 60572 60571 60570 60569 60568 60567 ...
 $ GAME_DATE        : chr  "2017-10-01" "2017-10-01" "2017-10-01" "2017-10-01" ...
 $ VIDEO            : chr  "https://www.mlb.com/video/presleys-solo-home-run/c-1858930483" "h
 $ PATH             : chr  "http://www.hittrackeronline.com/hrdetail.php?id=2017_6090" "http:/
 $ BATTER           : chr  "Alex Presley" "Rob Brantly" "Eric Hosmer" "Eric Young Jr." ...
 $ BATTER_TEAM      : chr  "DET" "CHW" "KC" "LAA" ...
 $ PITCHER          : chr  "Bartolo Colon" "Josh Tomlin" "Robbie Ray" "James Pazos" ...
 $ PITCHER_TEAM     : chr  "MIN" "CLE" "ARI" "SEA" ...
 $ INNING           : int  3 5 1 7 3 4 8 7 7 7 ...
 $ TYPE_LUCK        : chr  "JE" "JE/L" "JE/L" "JE" ...
 $ TRUE_DISTANCE    : int  343 346 353 367 376 381 383 384 387 392 ...
 $ EXIT_VELOCITY    : num  97.5 96.8 96.9 97.8 101.1 ...
 $ ELEVATION_ANGLE  : num  37.5 27.4 29.2 33.9 35.8 24 26.6 26.7 25.9 34.1 ...
 $ HORIZONTAL_ANGLE: num  61.7 55.7 118.7 112.2 113.2 ...
 $ APEX             : int  121 60 75 107 125 60 87 74 69 113 ...
 $ NUM_OF_PARKS     : int  28 3 27 27 29 30 30 30 30 30 ...
 $ BALLPARK         : chr  "Target Field" "Progressive Field" "Kauffman Stadium" "Angel Stadi
 $ X                : num  163 195 -170 -139 -148 ...
 $ Y                : num  302 286 310 340 346 ...
```

```
# data cleaning

# step 1
hr.step.1 <- hr.df.read %>%
  select(-RECORDID, -VIDEO, -PATH) %>% # drop these columns
  rename_all(tolower) %>% # make all columns lowercase
  rename(launch_angle = elevation_angle) # rename new = old
str(hr.step.1) # updated structure
```

```
'data.frame':   60576 obs. of  16 variables:
 $ game_date       : chr  "2017-10-01" "2017-10-01" "2017-10-01" "2017-10-01" ...
 $ batter          : chr  "Alex Presley" "Rob Brantly" "Eric Hosmer" "Eric Young Jr." ...
 $ batter_team     : chr  "DET" "CHW" "KC" "LAA" ...
 $ pitcher         : chr  "Bartolo Colon" "Josh Tomlin" "Robbie Ray" "James Pazos" ...
 $ pitcher_team    : chr  "MIN" "CLE" "ARI" "SEA" ...
 $ inning          : int  3 5 1 7 3 4 8 7 7 7 ...
 $ type_luck       : chr  "JE" "JE/L" "JE/L" "JE" ...
 $ true_distance   : int  343 346 353 367 376 381 383 384 387 392 ...
 $ exit_velocity   : num  97.5 96.8 96.9 97.8 101.1 ...
 $ launch_angle    : num  37.5 27.4 29.2 33.9 35.8 24 26.6 26.7 25.9 34.1 ...
 $ horizontal_angle: num  61.7 55.7 118.7 112.2 113.2 ...
 $ apex            : int  121 60 75 107 125 60 87 74 69 113 ...
 $ num_of_parks    : int  28 3 27 27 29 30 30 30 30 30 ...
 $ ballpark        : chr  "Target Field" "Progressive Field" "Kauffman Stadium" "Angel Stadi
 $ x               : num  163 195 -170 -139 -148 ...
 $ y               : num  302 286 310 340 346 ...
```

```
# step 2
hr.step.2 <- hr.step.1 %>%
    # creating new column, contact_type - case_when will create
    # categories within this column based on above launch angle
    # parameters.
mutate(contact_type = case_when(launch_angle < 10 ~ "Ground Ball",
    launch_angle >= 10 & launch_angle <= 25 ~ "Line Drive", launch_angle >
        25 & launch_angle <= 50 ~ "Fly Ball", launch_angle > 50 ~
        "Pop Up"))

# example of how contact_type works
head(hr.step.2[c("launch_angle", "contact_type")])
```

```
  launch_angle contact_type
1         37.5     Fly Ball
2         27.4     Fly Ball
3         29.2     Fly Ball
4         33.9     Fly Ball
5         35.8     Fly Ball
6         24.0   Line Drive
```

```
# making new frame that just includes ground ball home runs
ground.balls <- hr.step.2 %>%
    filter(contact_type == "Ground Ball")
```

```
# exit velocity and launch angle of the ground ball home runs
head(ground.balls[c("exit_velocity", "launch_angle")])
```

```
  exit_velocity launch_angle
1             0            0
2             0            0
3             0            0
4             0            0
5             0            0
6             0            0
```

```
# edited ground ball frame that just includes non 0 instances of
# launch angle (measurement error)
ground.balls.2 <- hr.step.2 %>%
    filter(contact_type == "Ground Ball" & launch_angle > 0)

# View
str(ground.balls.2)
```

```
'data.frame':   1 obs. of  17 variables:
 $ game_date       : chr "2011-05-28"
 $ batter          : chr "Carlos Gomez"
 $ batter_team     : chr "MIL"
 $ pitcher         : chr "Jonathan Sanchez"
 $ pitcher_team    : chr "SF"
 $ inning          : int 1
 $ type_luck       : chr "ITP/L"
 $ true_distance   : int 53
 $ exit_velocity   : num 76.9
 $ launch_angle    : num 0.8
 $ horizontal_angle: num 55.5
 $ apex            : int 3
 $ num_of_parks    : int 0
 $ ballpark        : chr "Miller Park"
 $ x               : num 30
 $ y               : num 43.7
 $ contact_type    : chr "Ground Ball"
```

```
# Need to make filtered data frame where all of our metrics have
# values that aren't zero
hr.step.3 <- hr.step.2 %>%
    filter(exit_velocity != 0 & launch_angle != 0 & apex != 0 & true_distance !=
        0 & horizontal_angle != 0)

# now the minimum exit velocity is 69.9 mph
summary(hr.step.3$exit_velocity)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   69.9   100.4   103.3   103.6   106.6   123.9 
```

```
# create stadium friendliness variable
stadium.friendliness <- hr.step.3 %>%
  group_by(ballpark) %>%
  summarize(total_home_runs = n()) %>% # will assign total home runs per ballpark
  ungroup() %>% # ungroup for later calculation purposes
  mutate(stadium_friendliness_measure = case_when(
    # case_when creates new measure with categories below
    total_home_runs > quantile(total_home_runs, 0.75) ~ "Very Friendly",
    total_home_runs > quantile(total_home_runs, 0.5) &
      total_home_runs <= quantile(total_home_runs, 0.75) ~ "Friendly",
    total_home_runs > quantile(total_home_runs, 0.25) &
      total_home_runs <= quantile(total_home_runs, 0.5) ~ "Not Friendly",
    total_home_runs <= quantile(total_home_runs, 0.25) ~ "Unfriendly"
  ))

# quick view
head(stadium.friendliness)
```

```
# A tibble: 6 x 3
  ballpark         total_home_runs stadium_friendliness_me~1
  <chr>                      <int> <chr>                    
1 AT&T Park                   1388 Not Friendly             
2 Ameriquest Field             179 Unfriendly               
3 Angel Stadium               1897 Friendly                 
4 BB&T Ballpark                  2 Unfriendly               
5 Busch Stadium               1730 Friendly                 
6 Chase Field                 2143 Very Friendly            
# i abbreviated name: 1: stadium_friendliness_measure
```

```
# final clean dataset
hr.clean <- hr.step.3
```

Data Visualizations and Tables

Figure 1. Launch Angle For Home Runs

```
# Create histogram with ggplot
ggplot(hr.clean, aes(x = launch_angle)) + geom_histogram(binwidth = 5,
    fill = "lightblue", color = "black") + labs(title = "Histogram of Launch Angle for Home
    x = "Launch Angle (degrees)", y = "Frequency") + scale_x_continuous(breaks = seq(0,
    50, by = 5)) + theme_minimal()
```

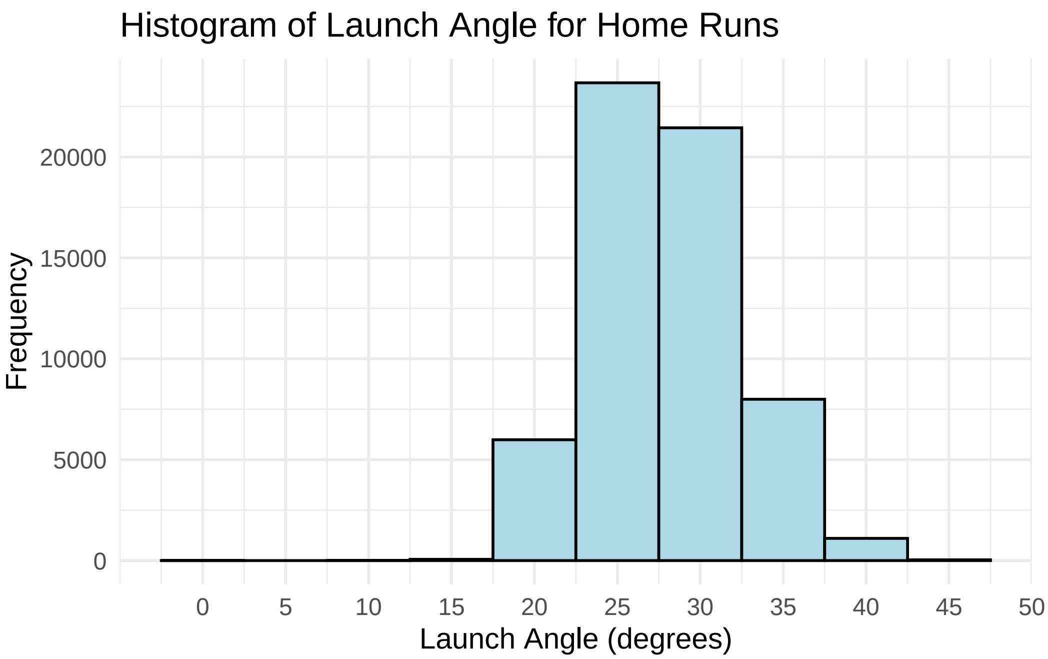

Table 1. Contact Type For Home Runs

```
# frequency counts for type of contact on home run balls, and
# associated launch angles
contact_la <- hr.clean %>%
    select(launch_angle, contact_type) %>%
    group_by(contact_type) %>%
    summarise(count = n(), mean.launch.angle = mean(launch_angle))
```

```r
# kbl table
kbl(contact_la, booktabs = TRUE) %>%
    kable_styling(latex_options = c("striped", "HOLD_position"))
```

contact_type	count	mean.launch.angle
Fly Ball	43908	29.82670
Ground Ball	1	0.80000
Line Drive	16403	22.87441

Figure 2. Exit Velocity

```r
# create variables of 90th percentile for exit velocity and
# distance - represented by dashed line
exit_velocity_90 <- quantile(hr.clean$exit_velocity, 0.9)
true_distance_90 <- quantile(hr.clean$true_distance, 0.9)

# Scatter plot of exit velocity and distance
ggplot(hr.clean, aes(x = exit_velocity, y = true_distance)) + geom_point() +
    labs(title = "Exit Velocity vs. Distance", x = "Exit Velocity (MPH)",
        y = "Distance (Feet)") + geom_hline(yintercept = true_distance_90,
    linetype = "dashed", color = "red") + geom_vline(xintercept = exit_velocity_90,
    linetype = "dashed", color = "red")
```

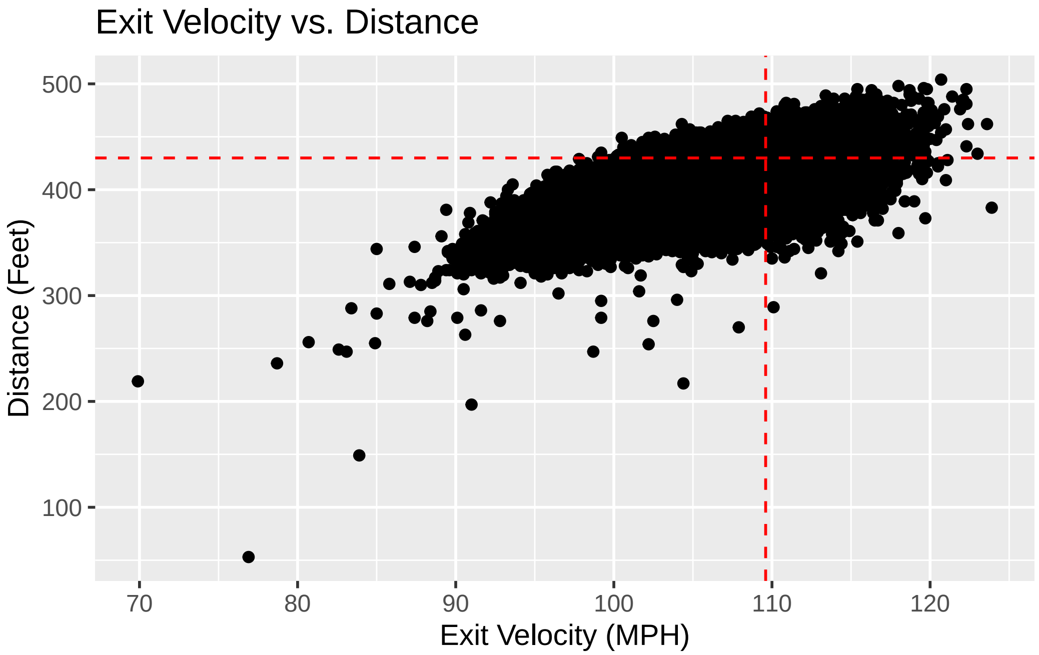

Table 2. Exit Velocity

```
# frequency counts for exit velocity intervals and associated average home run distance
ev_td <- hr.clean %>%
  mutate(ev_interval = cut(exit_velocity, breaks = seq(60, 150, by = 5), right = FALSE)) %>%
  select(ev_interval, true_distance) %>%
  group_by(ev_interval) %>%
  summarise(count = n(),
            mean.distance = mean(true_distance) # average home run distance
            )

# table
kbl(ev_td,
      col.names = c("Exit Velocity Interval", "Count", "Avg. Distance"),
      escape = FALSE) %>%
  kable_styling(latex_options= c("striped", "HOLD_position"))
```

Exit Velocity Interval	Count	Avg. Distance
[65,70)	1	219.0000
[75,80)	2	144.5000
[80,85)	6	240.6667
[85,90)	23	323.2609
[90,95)	1214	349.8451
[95,100)	12072	374.3996
[100,105)	24964	396.8924
[105,110)	16659	411.4811
[110,115)	4803	424.1226
[115,120)	543	438.7753
[120,125)	25	457.5200

Figure 3. Stadium Friendliness

```r
# create mean variable to add to boxplot
mean_num_of_parks <- mean(hr.clean$num_of_parks)

# Create boxplot with ggplot
ggplot(hr.clean, aes(x = "", y = num_of_parks)) + geom_boxplot(fill = "green",
    color = "black") + geom_hline(yintercept = mean_num_of_parks,
    linetype = "dashed", color = "red") + labs(title = "Home Run in How Many Ballparks?",
    x = NULL, y = "Ballparks") + theme_minimal()
```

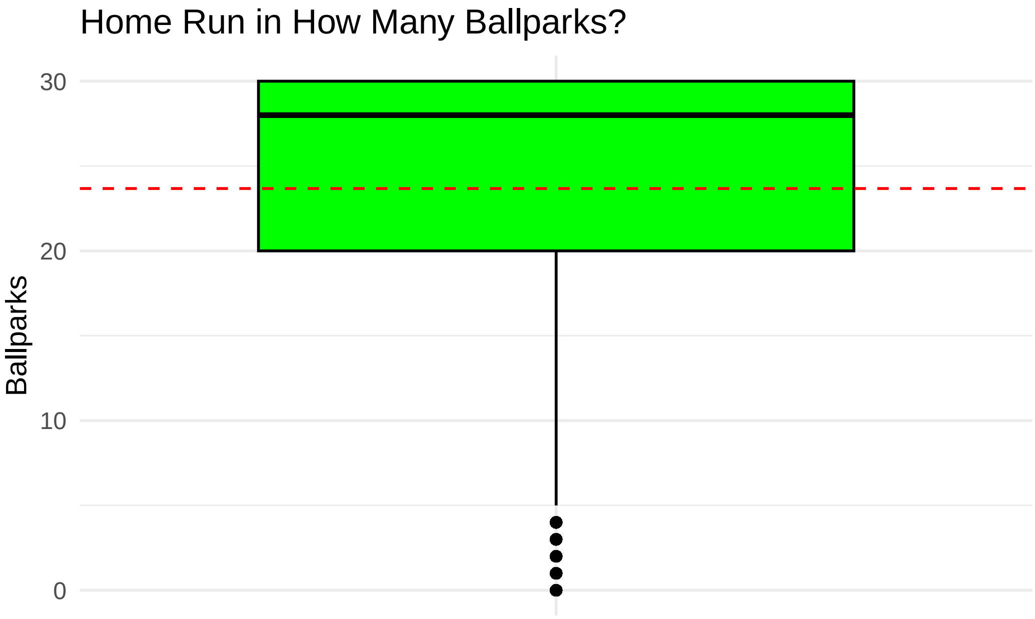

Table 3. Stadium Friendliness

```
# Arrange the data frame in descending order of total home run count
stadium_friendliness_sorted <- stadium.friendliness %>%
  filter(total_home_runs > 10) %>% # to avoid special games being played at stadiums rarely
  arrange(desc(total_home_runs)) # order from most to least

# Select the top 5 and bottom 5 stadiums
top_5_stadiums <- stadium_friendliness_sorted %>% slice_head(n = 5)
bottom_5_stadiums <- stadium_friendliness_sorted %>% slice_tail(n = 5)

# Combine top 5 and bottom 5 stadiums into one data frame
top_and_bottom_stadiums <- bind_rows(top_5_stadiums, bottom_5_stadiums)

# kbl table
kable(top_and_bottom_stadiums,
      col.names = c("Ballpark", "Home Run Count", "Stadium Friendliness"),
      escape = FALSE) %>%
  kable_styling(latex_options = c("striped", "HOLD_position"))
```

Ballpark	Home Run Count	Stadium Friendliness
Oriole Park at Camden Yards	2518	Very Friendly
Great American Ballpark	2466	Very Friendly
Rogers Centre	2365	Very Friendly
Miller Park	2342	Very Friendly
Citizens Bank Park	2334	Very Friendly
Sun Life Stadium	260	Unfriendly
Guaranteed Rate Field	229	Unfriendly
Ameriquest Field	179	Unfriendly
SunTrust Park	174	Unfriendly
Land Shark Stadium	169	Unfriendly

Figure 4. Temperature

```
# Read in weather dataset, and prep it to merge with home run
# dataset

# Rename venue_name to ballpark in baseball_weather
baseball_weather <- baseballr::get_game_info_sup_petti() %>%
    rename(ballpark = venue_name)

# Check data types
str(hr.clean$game_date)
```

```
 chr [1:60312] "2017-10-01" "2017-10-01" "2017-10-01" ...
```

```
str(baseball_weather$game_date)
```

```
 IDate[1:41946], format: "2022-10-05" "2022-10-05" "2022-10-05" ...
```

```
# Different data types on game date. Some ballparks in different
# capitalization

# Normalize game_date and ballpark data types between data sets
# so we can merge.

# Convert game_date to Date format and normalize ballpark in
```

```r
# hr.clean. To lower matched roughly another 1000 ballparks
hr.norm <- hr.clean %>%
    mutate(game_date = as.Date(game_date, format = "%Y-%m-%d"), ballpark = tolower(as.charac

# Convert game_date to Date format and normalize ballpark in
# baseball_weather
baseball_weather.norm <- baseball_weather %>%
    mutate(game_date = as.Date(game_date, format = "%Y-%m-%d"), ballpark = tolower(as.charac

# Merge on game_date and ballpark
hr.merged <- merge(hr.norm, baseball_weather.norm, by = c("game_date",
    "ballpark"), all.x = TRUE)

# Remove NA columns that don't match; dates have different
# ranges
hr.clean.merge <- na.omit(hr.merged)

str(hr.clean.merge)
```

```
'data.frame':	43561 obs. of  33 variables:
 $ game_date       : Date, format: "2008-03-30" ...
 $ ballpark        : chr  "nationals park" "nationals park" "citizens bank park" "citizens ba
 $ batter          : chr  "Chipper Jones" "Ryan Zimmerman" "Lastings Milledge" "Chase Utley"
 $ batter_team     : chr  "ATL" "WSH" "WSH" "PHI" ...
 $ pitcher         : chr  "Odalis Perez" "Peter Moylan" "Ryan Madson" "Matt Chico" ...
 $ pitcher_team    : chr  "WSH" "ATL" "PHI" "WSH" ...
 $ inning          : int  4 9 6 6 7 6 5 8 1 4 ...
 $ type_luck       : chr  "JE" "JE/L" "ND" "PL" ...
 $ true_distance   : int  413 407 430 365 408 399 381 403 418 371 ...
 $ exit_velocity   : num  109.7 110.9 107.3 94.8 100.3 ...
 $ launch_angle    : num  19.1 17.6 27.5 36.6 28.9 25.8 24.3 28.2 22.8 27.7 ...
 $ horizontal_angle: num  93.6 93.2 106.7 64.3 101.6 ...
 $ apex            : int  58 52 90 104 86 84 73 113 73 76 ...
 $ num_of_parks    : int  17 10 30 13 14 30 21 30 30 29 ...
 $ x               : num  -25.9 -22.7 -123.6 158.3 -82 ...
 $ y               : num  412 406 412 329 400 ...
 $ contact_type    : chr  "Line Drive" "Line Drive" "Fly Ball" "Fly Ball" ...
 $ game_pk         : int  233759 233759 233773 233773 233773 233765 233765 233765 233769 2337
 $ venue_id        : int  3309 3309 2681 2681 2681 2394 2394 2394 22 20 ...
 $ temperature     : int  49 49 51 51 51 51 51 51 62 80 ...
 $ other_weather   : chr  "Cloudy" "Cloudy" "Drizzle" "Drizzle" ...
 $ wind            : chr  "6 mph, Out To CF" "6 mph, Out To CF" "4 mph, R To L" "4 mph, R To
 $ attendance      : chr  "39,434" "39,434" "44,553" "44,553" ...
 $ start_time      : chr  "20:21:00" "20:21:00" "15:07:00" "15:07:00" ...
 $ elapsed_time    : chr  "2:22" "2:22" "2:56" "2:56" ...
 $ game_id         : chr  "2008/03/30/atlmlb-wasmlb-1" "2008/03/30/atlmlb-wasmlb-1" "2008/03/
 $ game_type       : chr  "R" "R" "R" "R" ...
 $ home_sport_code : chr  "mlb" "mlb" "mlb" "mlb" ...
 $ official_scorer : chr  "David Vincent" "David Vincent" "Mike Maconi" "Mike Maconi" ...
 $ date            : chr  "Att" "Att" "Att" "Att" ...
 $ status_ind      : chr  "F" "F" "F" "F" ...
 $ home_league_id  : int  104 104 104 104 104 103 103 103 104 104 ...
 $ gameday_sw      : chr  "E" "E" "E" "E" ...
 - attr(*, "na.action")= 'omit' Named int [1:17636] 1 2 3 4 5 6 7 8 9 10 ...
  ..- attr(*, "names")= chr [1:17636] "1" "2" "3" "4" ...
```

```
# home run proportions for every temperature

temperature_counts <- prop.table(table(hr.clean.merge$temperature))

# overall games played proportions for every temperature

temperature_counts2 <- prop.table(table(baseball_weather.norm$temperature))

# turn into data frames
temperature_counts <- data.frame(temperature_counts)
temperature_counts2 <- data.frame(temperature_counts2)

# Add group indicator so that we can combine and compare
temperature_counts$Group <- "All Home Runs"
temperature_counts2$Group <- "All Games"


# Combine data
combined_data <- rbind(temperature_counts, temperature_counts2)

combined_data$Var1 <- as.numeric(as.character(combined_data$Var1))

# Create plot
ggplot(combined_data, aes(x = Var1, y = Freq, fill = Group)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Comparison of Proportion Tables",
       x = "Temperature (Degrees Fahrenheit)",
       y = "Proportion") +
  scale_x_continuous(breaks = seq(min(combined_data$Var1), max(combined_data$Var1), by = 10)
  theme_minimal()
```

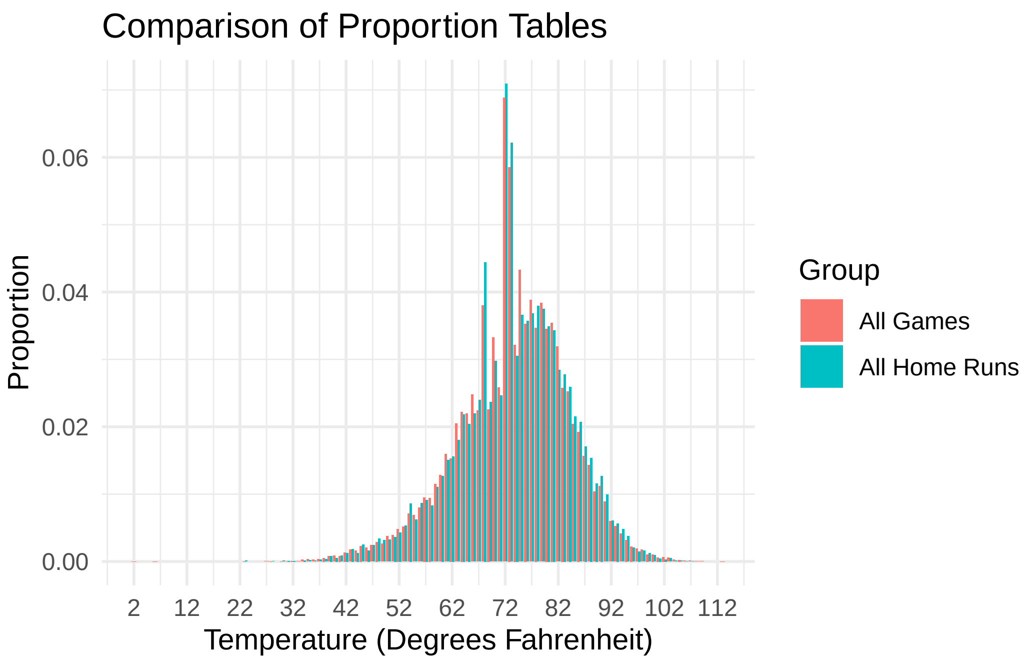

Table 4. Temperature

```
# Pivot the data into wider format
combined_data_wide <- combined_data %>%
  pivot_wider(names_from = Group, values_from = Freq) %>% # To view both side by side
  mutate(`All Home Runs` = `All Home Runs` * 100) %>%
  mutate(`All Games` = `All Games` * 100) %>% # get into more readable format for comparison
  mutate(Difference = `All Home Runs` - `All Games`) %>% # Difference between groups
  filter(Var1 >= 65 & Var1 <= 75) # filter to bulk of temperature games are played at

kable(combined_data_wide, caption = "Combined Temperature Percentages",
      col.names = c("Temperature", "All Home Runs", "All Games", "Difference"),
      escape = FALSE) %>%
  kable_styling(latex_options = c("striped", "HOLD_position"))
```

Table 1: Combined Temperature Percentages

Temperature	All Home Runs	All Games	Difference
65	2.043112	2.198064	-0.1549522
66	2.196919	2.476994	-0.2800750
67	2.396639	2.243361	0.1532787
68	4.444342	3.802508	0.6418344
69	2.366796	2.260049	0.1067474
70	2.975138	3.323320	-0.3481821
71	2.463213	2.584275	-0.1210625
72	7.093501	6.885043	0.2084584
73	6.216570	5.855147	0.3614228
74	3.055485	3.218424	-0.1629383
75	3.661532	4.329376	-0.6678438

```r
# 2-Sample-Proportion Test to see if the home run proportions at
# 68F is smaller Ho: p1-p2 <= 0; Ha: p1-p2 > 0

n1 <- sum(table(hr.clean.merge$temperature))
n2 <- sum(table(baseball_weather.norm$temperature))

x1 <- sum(hr.clean.merge$temperature == 68)
x2 <- sum(baseball_weather.norm$temperature == 68)

# normality check:
phat1 <- x1/n1
phat2 <- x2/n2

print(phat1 * (1 - phat1) * n1)
```

```
[1] 1849.958
```

```r
print(phat2 * (1 - phat2) * n2)
```

```
[1] 1534.35
```

```r
prop.test(x = c(x1, x2), n = c(n1, n2), alternative = "greater")
```

```
    2-sample test for equality of proportions with
    continuity correction

data:  c(x1, x2) out of c(n1, n2)
X-squared = 22.074, df = 1, p-value = 0.000001312
alternative hypothesis: greater
95 percent confidence interval:
 0.004159538 1.000000000
sample estimates:
    prop 1     prop 2 
0.04444342 0.03802508 
```

Both normality checks passed as they are greater than or equal to 10. We can then use z and proceed with our test. The p-value is smaller than any alpha value used in practice, so we reject our null hypothesis that the home run proportion is smaller. Thus, a temperature of 68F does have a statistical difference in terms of home run potential - It is more conducive to home run hitting.

Figure 5. Type of Day

```r
# Calculate the proportion of each day type category in
# hr.clean.merge dataset
prop_hr_daytype <- prop.table(table(hr.clean.merge$other_weather)) *
    100

# Get the top day types with proportions for hr.clean.merge
# dataset
top_hr_daytype <- sort(prop_hr_daytype, decreasing = TRUE)[1:11]

# Calculate the proportion of each day type category in weather
# dataset
prop_all_daytype <- prop.table(table(baseball_weather.norm$other_weather)) *
    100

# Get the top 5 day types with proportions for weather dataset
top_all_daytype <- sort(prop_all_daytype, decreasing = TRUE)[1:11]


# make into data frame
top_hr_daytype <- data.frame(top_hr_daytype)
```

```r
top_all_daytype <- data.frame(top_all_daytype)

# Add group indicator
top_hr_daytype$Group <- "Home Run %"
top_all_daytype$Group <- "Weather Occurrence %"

combined_data_daytype <- rbind(top_hr_daytype, top_all_daytype)


library(ggplot2)
ggplot(combined_data_daytype, aes(x = Var1, y = Freq, fill = Group)) +
    geom_bar(stat = "identity", position = "dodge") + labs(title = "Impact of the Type of Day
    x = "Category", y = "Percentage") + theme_minimal() + theme(axis.text.x = element_text(si
```

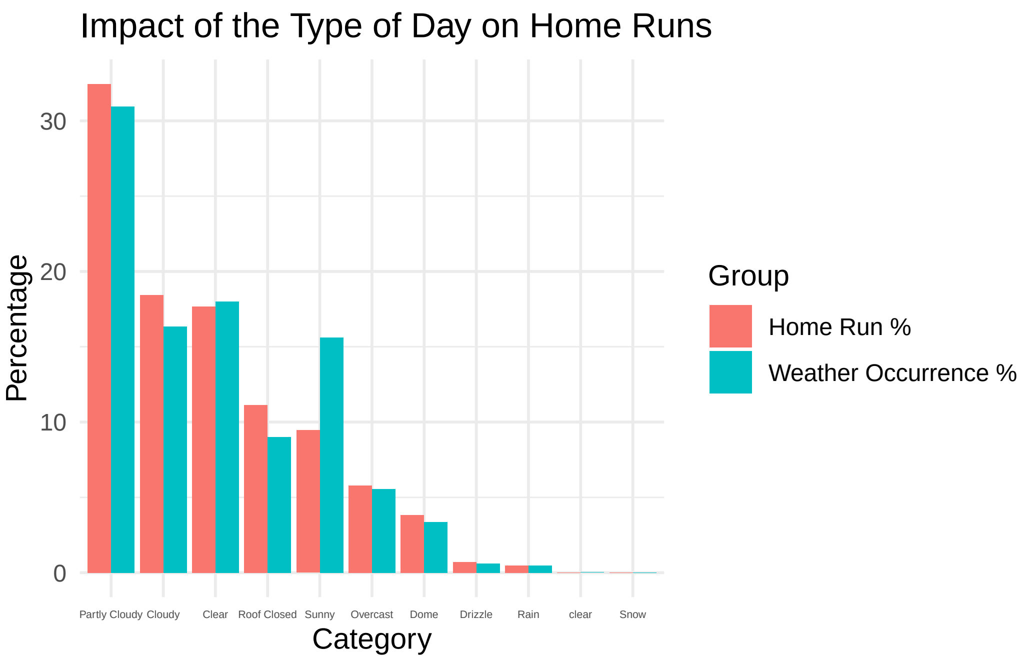

Table 5. Type of Day

```r
# Pivot the data into wider format
combined_data_wide_daytype <- combined_data_daytype %>%
  pivot_wider(names_from = Group, values_from = Freq) %>% # To view both side by side
  mutate(Difference = `Home Run %` - `Weather Occurrence %`)
```

```
kable(combined_data_wide_daytype, caption = "Combined Day Type Percentages",
      col.names = c("Day Type", "Home Run %", "Weather Occurrence %", "Difference"),
      escape = FALSE) %>%
  kable_styling(latex_options = c("striped", "HOLD_position"))
```

Table 2: Combined Day Type Percentages

Day Type	Home Run %	Weather Occurrence %	Difference
Partly Cloudy	32.4533413	30.9588519	1.4944894
Cloudy	18.4362159	16.3329042	2.1033117
Clear	17.6809531	18.0088685	-0.3279154
Roof Closed	11.1315167	9.0139703	2.1175464
Sunny	9.4557058	15.6057789	-6.1500731
Overcast	5.7781043	5.5642970	0.2138073
Dome	3.8337045	3.3757688	0.4579356
Drizzle	0.7162370	0.6150765	0.1011605
Rain	0.4866739	0.4839556	0.0027183
clear	0.0252519	0.0309922	-0.0057403
Snow	0.0022956	0.0095361	-0.0072404

```
# 2-Sample-Proportion Test to see if the home run proportion on
# a Sunny day is smaller Ho: p1-p2 >= 0; Ha: p1-p2 < 0

n1 <- sum(table(hr.clean.merge$other_weather))
n2 <- sum(table(baseball_weather.norm$other_weather))

x1 <- sum(hr.clean.merge$other_weather == "Sunny")
x2 <- sum(baseball_weather.norm$other_weather == "Sunny")

# normality check:
phat1 <- x1/n1
phat2 <- x2/n2

print(phat1 * (1 - phat1) * n1)
```

```
[1] 3729.519
```

```r
print(phat2 * (1 - phat2) * n2)
```

```
[1] 5524.446
```

```r
prop.test(x = c(x1, x2), n = c(n1, n2), alternative = "less")
```

```

    2-sample test for equality of proportions with
    continuity correction

data:  c(x1, x2) out of c(n1, n2)
X-squared = 739.8, df = 1, p-value <
0.00000000000000022
alternative hypothesis: less
95 percent confidence interval:
 -1.00000000 -0.05776081
sample estimates:
    prop 1     prop 2 
0.09455706 0.15605779 
```

Both normality checks passed as they are greater than or equal to 10. We can then use z and proceed with our test. The p-value is smaller than any alpha value used in practice, so we reject our null hypothesis that the the home run proportion is bigger. Thus, a sunny day does have a statistical difference in terms of home run potential - It is less conducive to home run hitting.

Conclusions

After conducting this analysis, I can now provide some insights to my original research questions:

- What is the optimal exit velocity and launch angle for hitting home runs?

The majority of home runs are hit at exit velocities between 100-105 mph, followed closely by the interval 105-110 mph. Launch angle is most dominant from 22.5-27.5 degrees, but plenty of home runs are hit up to 32.5 degrees and beyond. It was rare to come across home runs in our dataset with exit velocities above 115 mph. Is this because it is that much more difficult to hit a ball that hard, so there are significantly less instances of it - or does a higher exit velocity make it harder for a ball to have enough launch to get over the fence? This is where it would have been useful to have data on non-home run at bats as well to have a comparison of hard hit line drives that don't leave the ballpark.

- What stadiums are the most hitter friendly/have the most home runs - does weather have an impact?

The top 2 most hitter friendly parks in terms of total home runs hit from 2006-2017 are Oriole Park at Camden Yards and Great American Ballpark (Reds). However, plenty of new stadiums were being built during this time period, so an accurate stadium friendliness measure wasn't able to be calculated for them. It would be helpful to conduct this analysis again in a time period with less stadium turnover.

Weather does seem to have an impact on home runs. While I wasn't able to observe an overall pattern of temperature and home runs, certain temperature points are statistically shown to be more conducive to home run hitting, such as 68F. Given that temperature points around 68F don't consistently show the same results, it is difficult to determine if certain temperature ranges effect home run hitting more. This will need further study.

The type of day seems to have a more clear impact. Games played on sunny days have a statistically significant negative effect on home run hitting while games played with the roof closed have a statistically significant positive effect on home run hitting. It is interesting that sunny days seem to be a disadvantage to home run hitting, I am wondering if the glare of the sun may impact the vision of the batter at the plate. A closed roof being an advantage to home run hitting may need more investigation - only seven MLB teams have a retractable roof, so not everyone has the same chance of playing in this type of environment regularly. In other words, the few teams that have a retractable roof could have other home run advantages, such as having a smaller, more home run friendly ballpark.