What Makes an Effective Pitcher?

Anand Tanna & Riku Komatani

Introduction

There are many factors that define an effective pitcher in the MLB. Our mission is to explore these factors in order to determine what differentiates a good pitcher from a bad pitcher.

We will look at:

- Pitch zone location
- Pitch velocity + movement
- Pitch types + pitch combinations
- Important pitching statistics: SO, BB, WAR, etc.

Data

We will be using Statcast data from 2018-2022 (excluding 2020). Our data is scraped from Baseball Savant, Baseball Reference, and the pitch-by-pitch statcast data from Lab 4.

- Only included pitchers who pitched 50 innings or more in a single season

Most and Least Valuable Pitchers based off WAR

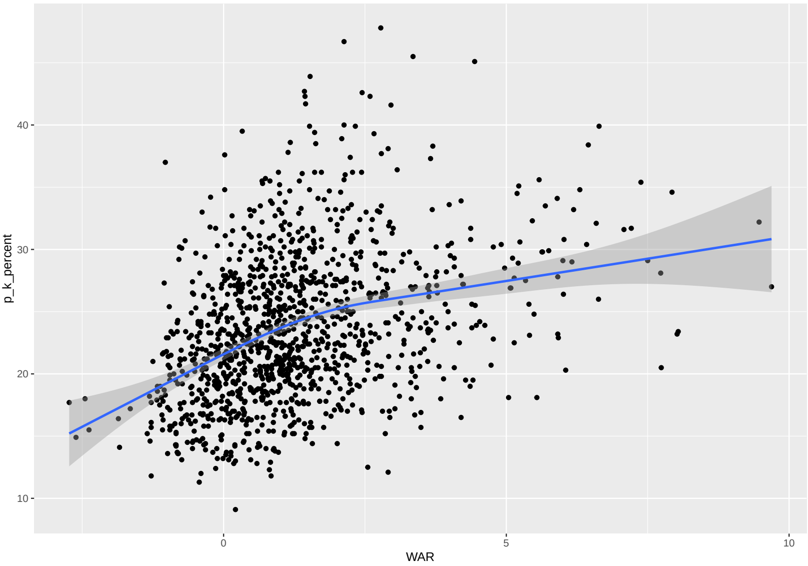
```
statcast_data2 %>%
  select(pitcher_name, year, player_age, p_formatted_ip, p_era, WAR) %>%
  arrange(desc(WAR)) %>% head(10)
```

```
##         pitcher_name year player_age p_formatted_ip p_era  WAR
## 1         Aaron Nola 2018         25          212.1  2.37 9.69
## 2       Jacob deGrom 2018         30          217.0  1.70 9.47
## 3    Sandy Alcantara 2022         26          228.2  2.28 8.04
## 4         Mike Minor 2019         31          208.1  3.59 8.02
## 5       Max Scherzer 2018         33          220.2  2.53 7.93
## 6      Kyle Freeland 2018         25          202.1  2.85 7.74
## 7         Lance Lynn 2019         32          208.1  3.67 7.73
## 8       Zack Wheeler 2021         31          213.1  2.78 7.50
## 9   Justin Verlander 2019         36          223.0  2.58 7.38
## 10      Jacob deGrom 2019         31          204.0  2.43 7.21
```

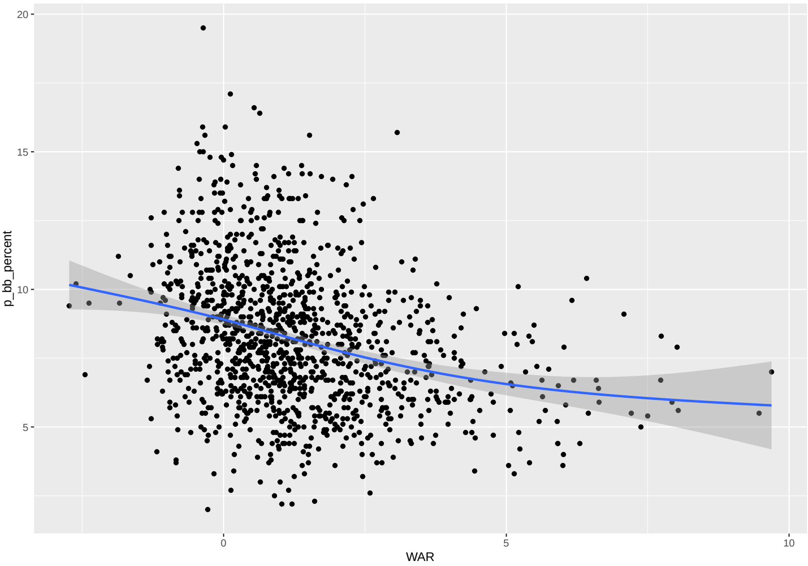
```
statcast_data2 %>%
  select(pitcher_name, year, player_age, p_formatted_ip, p_era, WAR) %>%
  arrange(desc(WAR)) %>% tail(10)
```

```
##           pitcher_name year player_age p_formatted_ip p_era   WAR
## 1210       Dylan Covey 2019         27           58.2  7.98 -1.30
## 1211         J.A. Happ 2021         38          152.1  5.79 -1.31
## 1212      Homer Bailey 2018         32          106.1  6.09 -1.35
## 1213     Brett de Geus 2021         23           50.0  7.56 -1.65
## 1214    Matt Shoemaker 2021         34           60.1  8.06 -1.84
## 1215 Justus Sheffield 2021         25           80.1  6.83 -1.86
## 1216     Edwin Jackson 2019         35           67.2  9.58 -2.38
## 1217    Patrick Corbin 2022         32          152.2  6.31 -2.45
## 1218    Dallas Keuchel 2022         34           60.2  9.20 -2.61
## 1219      Jake Arrieta 2021         35           98.2  7.39 -2.73
```

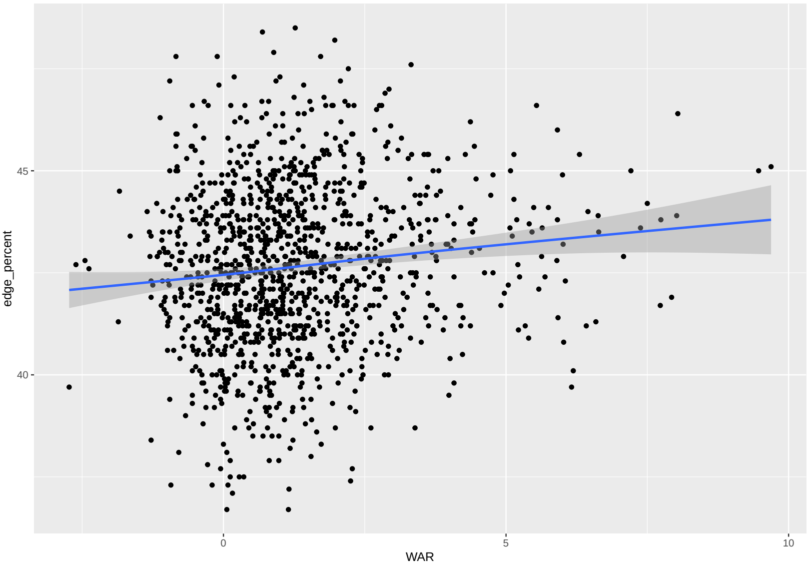
Strikeout % vs. WAR



Walk % vs. WAR



Edge % vs. WAR



Ohtani (2022) vs. Keller (2021)

Both of these pitchers throw at a high velocity, however Ohtani threw much more effectively. Here we look at their pitch types.

Ohtani:

```
## # A tibble: 7 x 3
##   pitch_type     N    pct
##   <chr>      <int>  <dbl>
## 1 SL           779 0.406
## 2 FF           457 0.238
## 3 FS           204 0.106
## 4 FC           189 0.0985
## 5 CU           155 0.0808
## 6 SI            80 0.0417
## 7 <NA>          55 0.0287
```

Keller:

```
## # A tibble: 4 x 3
##   pitch_type     N    pct
##   <chr>      <int>  <dbl>
## 1 FF          1608 0.571
## 2 SL           653 0.232
## 3 CU           428 0.152
## 4 CH           126 0.0448
```

Ohtani vs. Keller

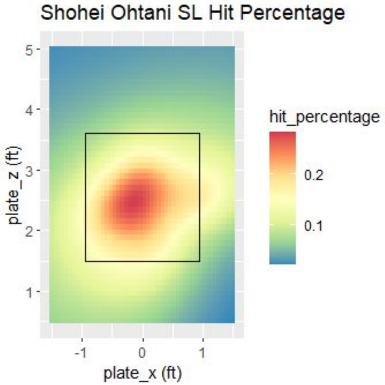
Here are stats that stand out the most, Ohtani makes batters whiff more and his slider has much more movement.

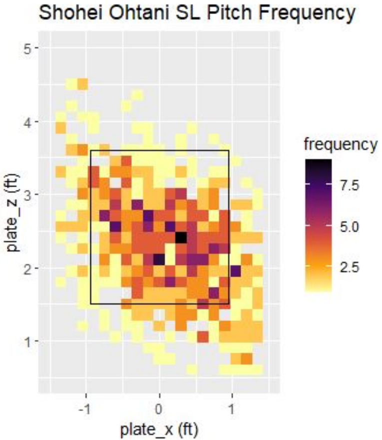
```
##    pitcher_name whiff_percent fastball_avg_speed sl_avg_speed sl_avg_spin
## 1 Shohei Ohtani            33               95.7         85.3        2492
##   sl_avg_break edge_percent p_bb_percent
## 1           15         40.1          6.7
```

```
##   pitcher_name whiff_percent fastball_avg_speed sl_avg_speed sl_avg_spin
## 1 Mitch Keller          20.2               93.8         86.1        2370
##   sl_avg_break edge_percent p_bb_percent
## 1          3.8         41.4         10.4
```

Ohtani Slider Pitch Location Chart

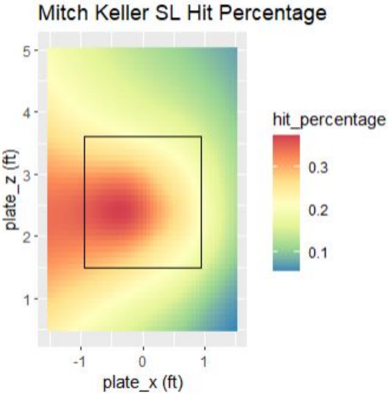
- Ohtani tends to throw his slider in various locations within the strike zone. He might choose to throw sliders in any location within the strike zone because it breaks a lot.

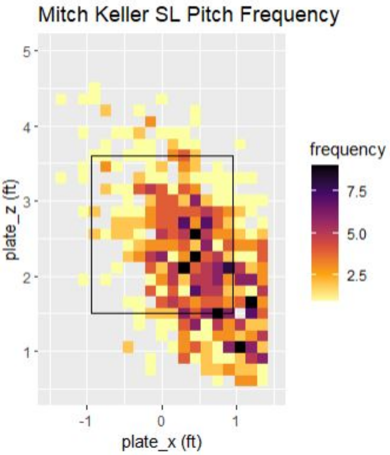




Keller Slider Pitch Location Chart

- Keller tends to throw sliders on outside corner (for righties). He has to throw his slider on the outside corner (for righties) due to its lack of movements.





Chapman vs. Rogers

Here we have 2 bullpen pitchers. They are both effective, yet their pitching styles are very different. Here we look at their pitch types.

Chapman:

```
## # A tibble: 5 x 3
##   pitch_type     N    pct
##   <chr>      <int>  <dbl>
## 1 FF          2643 0.640
## 2 SL          1078 0.261
## 3 SI           245 0.0594
## 4 FS           120 0.0291
## 5 CH            42 0.0102
```

Rogers:

```
## # A tibble: 4 x 3
##   pitch_type     N     pct
##   <chr>      <int>   <dbl>
## 1 FF           775 0.527
## 2 SL           601 0.409
## 3 SI            85 0.0578
## 4 CU            10 0.00680
```

Chapman vs. Rogers

Here are the stats that stand out the most. Chapman relies on high velocity, while Rogers throws relatively slow but still has a higher WAR than Chapman. He relies on groundball outs while Chapman relies on the batter whiffing. Rogers also has more break on his offspeed pitches. Rogers also has better command as he threw less walks in more innings pitched.

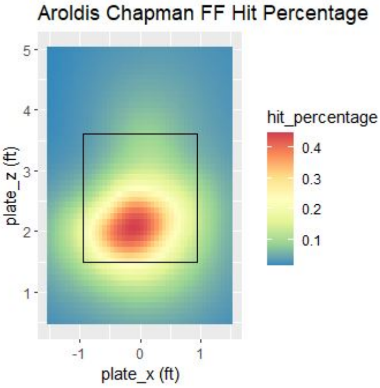
```
##      pitcher_name fastball_avg_speed fastball_avg_spin sl_avg_speed
## 1 Aroldis Chapman               98.3              2492         85.2
##   sl_avg_break whiff_percent p_walk groundballs_percent  WAR
## 1         11.2          31.5     25                42.3 1.61
```

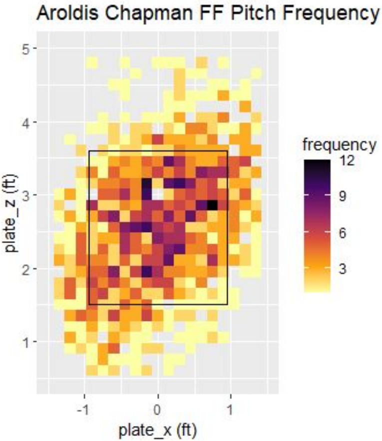
```
statcast_data2 %>%
  filter(pitcher_name == "Tyler Rogers", year == 2021) %>%
  select(pitcher_name, fastball_avg_speed, fastball_avg_spin, sl_avg_speed, sl_avg_break, whi
```

```
##   pitcher_name fastball_avg_speed fastball_avg_spin sl_avg_speed sl_avg_break
## 1 Tyler Rogers               82.7              1856         71.8         19.3
##   whiff_percent p_walk groundballs_percent  WAR
## 1          16.5     13                58.1 2.45
```

Chapman Fastball Pitch Location Chart

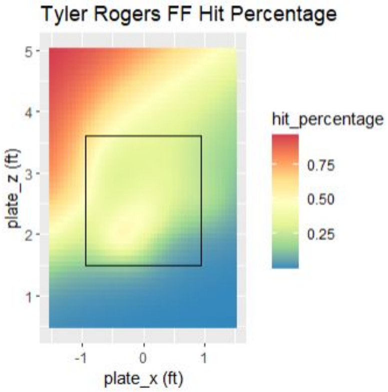
- Chapman throws his fastball from top to bottom of the strike zone because his high velocity makes it hard to hit in any location. Chapman's hit percentage is low in the higher part of the strike zone.

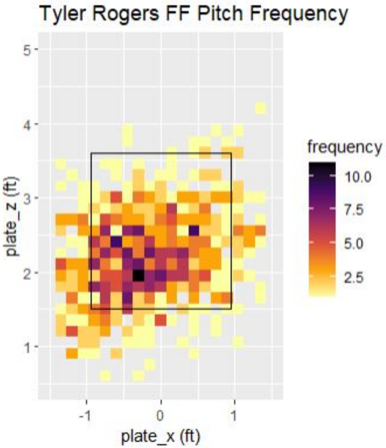




Rogers Fastball Pitch Location Chart

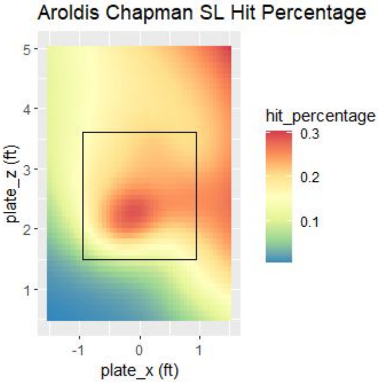
- Rogers tends to throw fastball in low strike zones to stay safe since he throws a slow fastball. Rogers hit percentage is around 0.3 - 0.5 for locations that he throws often, which is higher than expected.

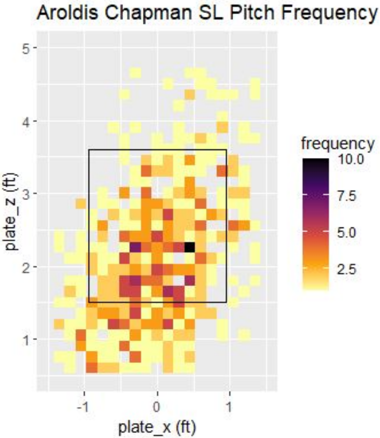




Chapman Slider Pitch Location Chart

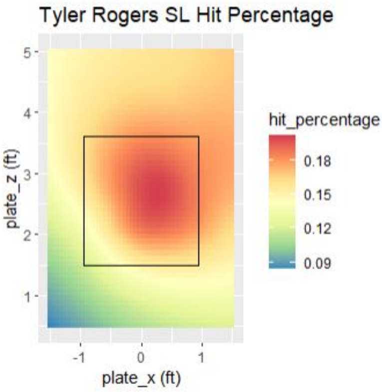
- Chapman tends to throw sliders on low inside corner (for righties) into the ball zone which tends to be most effective place to throw for lefty pitchers.

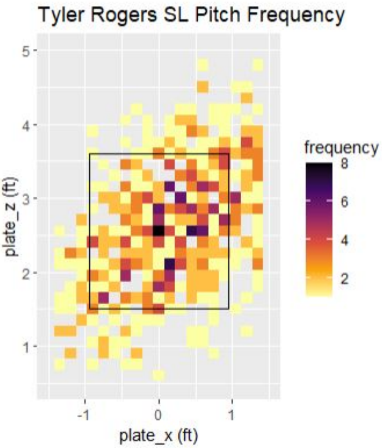




Rogers Slider Pitch Location Chart

- Rogers tends to throw sliders a little more often on high outside corner (for righties) which is reasonable since he is a right side-arm pitcher and his slider will tail away from righty batters in that location, making it hard to hit.





Best model for predicting WAR of pitchers

```r
set.seed(13)
ind <- sample(1:nrow(statcast_data2), size = 150, replace = FALSE)
train <- p_data3[ind, ]
test <- p_data3[-ind, ]

m_small <- lm(WAR ~ p_k_percent + p_bb_percent, data = train)
test1 <- test
test1$Prediction <- predict(m_small, test1)
sqrt(mean((test1$WAR - test1$Prediction)^2))
```

```
## [1] 1.437425
```

```r
m_big <- lm(WAR ~ p_k_percent + p_bb_percent + in_zone_percent + whiff_percent +
f_strike_percent + fastball_avg_break + fastball_avg_speed + offspeed_avg_speed
+ offspeed_avg_break +  meatball_percent + groundballs_percent +
flyballs_percent + edge_percent, data = train)
test2 <- test
test2$Prediction <- predict(m_big, test2)
sqrt(mean((test2$WAR - test2$Prediction)^2))
```

```
## [1] 1.451718
```

```r
m_best <- lm(WAR ~ p_k_percent + p_bb_percent + whiff_percent +
meatball_percent + groundballs_percent + fastball_avg_spin , data = train)
test3 <- test
test3$Prediction <- predict(m_best, test3)
sqrt(mean((test3$WAR - test3$Prediction)^2))
```

```
## [1] 1.399853
```

Best model for predicting WAR of pitchers

```
summary(m_best)
```

```
##
## Call:
## lm(formula = WAR ~ p_k_percent + p_bb_percent + whiff_percent +
##     meatball_percent + groundballs_percent + fastball_avg_spin,
##     data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6831 -0.8308 -0.2281  0.4855  5.3656
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -3.4080717  2.5428437  -1.340 0.182554
## p_k_percent          0.1319585  0.0476845   2.767 0.006497 **
## p_bb_percent        -0.2147722  0.0555644  -3.865 0.000176 ***
## whiff_percent       -0.0220969  0.0504145  -0.438 0.661910
## meatball_percent    -0.1068780  0.1064620  -1.004 0.317332
## groundballs_percent  0.0309362  0.0180219   1.717 0.088492 .
## fastball_avg_spin    0.0015116  0.0009803   1.542 0.125563
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.448 on 127 degrees of freedom
##   (16 observations deleted due to missingness)
## Multiple R-squared:  0.3023, Adjusted R-squared:  0.2693
## F-statistic: 9.169 on 6 and 127 DF,  p-value: 2.461e-08
```

Conclusions

- Our model found that strikeout %, walk %, whiff %, meatball %, groundball %, and fastball spin were the best predicting variables for calculating the value of a pitcher.
- There is no set way for a pitcher to be effective. Through our comparisons, we saw it takes a combination of many factors such as pitch locations and command, pitch type usage, velocity, throwing mechanics, and pitch movement to be an effective MLB pitcher.
- It is important to look at factors together and not individually. For example, edge % is better to look at with pitch movement rather than by it self.

Code for Hit Percentage Heat Plot

```
#Function for hit percentage heat plot
hit_likely <- function(data) {
  data <- data %>%
    mutate(Hit = ifelse(events %in% c("single", "double", "triple", "home_run"), 1, 0))
  # implement the GAM fit binary model (logistic link)
  fit <- gam(Hit ~ s(plate_x, plate_z), family = binomial, data = data)
  # find predicted probabilities over a 50 x 50 grid
  x <- seq(-1.5, 1.5, length.out=50)
  y <- seq(0.5, 5, length.out=50)
  data.predict <- data.frame(plate_x = c(outer(x, y * 0 + 1)),
                             plate_z = c(outer(x * 0 + 1, y)))
  predicted_data <- fit %>%
    augment(type.predict = "response", newdata = data.predict)
  colnames(predicted_data)[colnames(predicted_data) == ".fitted"] <- "hit_percentage"
  # construct heat percentage tile plot with strike zone boundary line
  ggplot(predicted_data, aes(plate_x, plate_z)) +
    geom_tile(aes(fill = hit_percentage)) +
    scale_fill_distiller(palette = "Spectral") +
    geom_path(data = strike_zone, aes(x, y)) +
    coord_fixed() + xlab("plate_x (ft)") + ylab("plate_z (ft)")
}
```

Code for Pitch Location Frequency Heat Plot

```r
#Function for pitch location frequency
pitch_freq <- function(data) {
  #Divide zone into box of 0.15 (ft^2)
  data <- data %>%
    mutate(plate_x_fifteenth = 0.15 * round(plate_x / 0.15, 0),
           plate_z_fifteenth = 0.15 * round(plate_z / 0.15, 0))
  #Count the frequency of pitch in each box
  pitch_grouped <- data %>%
    dplyr::group_by(plate_x_fifteenth, plate_z_fifteenth) %>%
    dplyr::summarize(frequency = n())
  #Plot pitch location frequency with strike zone boundary line
  ggplot(pitch_grouped, aes(plate_x_fifteenth, plate_z_fifteenth)) +
    geom_tile(aes(fill = frequency)) +
    geom_path(data = strike_zone, aes(x, y)) +
    scale_fill_viridis_c(option = "B", direction = -1) +
    coord_fixed() +
    xlim(c(-1.5, 1.5)) +
    ylim(c(0.5, 5)) + xlab("plate_x (ft)") + ylab("plate_z (ft)")
}
```

Shiny App

- Now it is time to demo our Shiny App where you will be able to see hit percentage and pitch location frequency of pitchers!
- This app allows you to select the pitcher, year, and pitch type.