Trout in the Statcast Era

Walter Limm

12/11/2021

https://baseballsavant.mlb.com/csv-docs (https://baseballsavant.mlb.com/csv-docs)

Introduction

In 2015, MLB installed the camera tracking system “Statcast” in all 30 stadiums. This system tracks a large amount of information, from basics like how hard a pitch was thrown or how hard it was hit to things like the pitcher’s release point or the spin axis of the pitch. My dataset is Statcast data accessed through baseballsavant.mlb.com. It includes data from every pitch thrown to the hitter Mike Trout since 2015. Mike Trout has been among the best hitters in baseball since his debut in 2011, winning multiple MVP awards and consistently being at the top of important hitting statistic leaderboards, and I’d like to investigate what makes him great and how a pitcher should approach him. To do this, I’m looking to answer three questions:

- What pitch locations does Mike Trout hit the best/worst?
- What pitch types does Mike trout hit the best/worst?
- Based on pitch location and type, how well could you expect Mike Trout to hit?

A few things need to be defined before diving into the data. The obvious questions that arise are “What constitutes good hitting?” or “what metrics will we use to distinguish best from worst?” In my opinion, there are many possible good options. You could use a myriad of hitting stats like batting average, on-base percentage, slugging percentage and more. However, all of those stats are based on outcomes of at-bats, not the outcomes of individual pitches. Because we are looking at data from individual pitches, I think we should use a custom metric.

I will judge pitches on a binary by dividing them into “good” and “bad” outcomes for Trout. Good outcomes will include pitches that are “Hard hit” which is defined by Statcast as a ball with an exit velocity of 95 mph or greater. Pitches that Trout took for a ball will also be considered good. Bad outcomes are pitches that Trout swung at and missed, hit with an exit velocity below 95 mph, and pitches that he takes for a strike. Pitches Trout fouls off will also be considered a bad outcome for him. I see a possible point of confusion from this metric. I’m using it to measure how good someone is at hitting, but also include instances where Trout makes the choice to not swing at pitches outside of the strike zone as a good result. Because he is not swinging, he is not hitting. Here’s the disctinction: I’m not looking at hitting as only the act of making contact with the ball - instead “hitting” is a proxy term for offensive production. It is good for offensive production for a batter to take pitches outside of the strike zone because it gets them closer to a walk, which is why it is sorted as a good result.

There is a secondary reason for defining a binary metric. Because this data is already tidy (In the factors I’m using), I want to include something beyond what we completed in class in the analysis section. By using a binary metric, I can use a logistic regression to see how well pitch location and type predict a good or bad result.

Setting up the Data

```
Trout <- read_csv("TroutInStatcast.csv")
```

```
## New names:
## * pitcher -> pitcher...8
## * fielder_2 -> fielder_2...42
## * pitcher -> pitcher...60
## * fielder_2 -> fielder_2...61
```

```
## Rows: 14806 Columns: 92
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (18): pitch_type, game_date, player_name, events, description, des, game...
## dbl (67): release_speed, release_pos_x, release_pos_z, batter, pitcher...8, ...
## lgl  (7): spin_dir, spin_rate_deprecated, break_angle_deprecated, break_leng...
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

There's a lot of pitches here - the dataset is 14806 observations of 92 factors. We don't care about a lot of factors - the interesting ones to us are: pitch_type (what kind of pitch was thrown), description (the outcome - ball, swing and miss, etc.), zone (where the pitch was thrown), and launch_speed (exit velocity). plate_x and plate_z are preserved for later graphing purposes. Let's pare down this set to those factors

```
Trout <- Trout %>% select(pitch_type, description, zone, launch_speed, plate_x, plate_z)
```

Let's look at the different kinds of outcomes.

```
Trout %>% count(description)
```

```
## # A tibble: 11 x 2
##    description                 n
##    <chr>                   <int>
##  1 ball                     5844
##  2 blocked_ball              326
##  3 called_strike            2923
##  4 foul                     2356
##  5 foul_tip                  106
##  6 hit_by_pitch               59
##  7 hit_into_play            2084
##  8 intent_ball                88
##  9 pitchout                    1
## 10 swinging_strike           918
## 11 swinging_strike_blocked   101
```

There are 11 different tracked outcomes. Of these, ball, blocked_ball, hit_by_pitch, intent_ball, and pitchout will all be good outcomes for Trout. called_strike, foul, foul_tip, swinging_strike, and swinging_strike_blocked are all bad outcomes. hit_into_play will depend on the launch_speed of the ball in play

To make the mutate easier, I will first make NA values in the launch_speed factor equal to 0 (instances where the ball wasn't hit).

```
Trout$launch_speed[is.na(Trout$launch_speed)] <- 0
```

This code defines sets the good outcomes that I've layed out as "1" in the new factor result.

```
Trout <- Trout %>% mutate(result = ifelse(launch_speed >= 95 & description == "hit_into_play" |
                                 description == "ball" |
                                 description == "blocked_ball" |
                                 description == "hit_by_pitch" |
                                 description == "intent_ball" |
                                 description == "pitchout",
                                 1, 0))
```

Analysis

Question 1: What pitch locations does Mike Trout hit the best/worst?

```
Trout %>% count(zone)
```

```
## # A tibble: 14 x 2
##     zone     n
##    <dbl> <int>
##  1     1   595
##  2     2   734
##  3     3   577
##  4     4   799
##  5     5   988
##  6     6  1005
##  7     7   568
##  8     8   829
##  9     9   929
## 10    11  1520
## 11    12  1472
## 12    13  1434
## 13    14  3213
## 14    NA   143
```

Trout saw pitches in 14 zones. 143 of the total pitches have an NA value as their zone. I've looked through the descriptions in the tibble below, but I can't find a discernable pattern. Unfortunately, the MLB players Union and MLB Owners are disputing after the expiration of the previous Collective Bargain Agreement, and the MLB is in lockout. As a result, all media involving current MLB players has been scrubbed from MLB websites, so I cannot find game footage to cross-check.

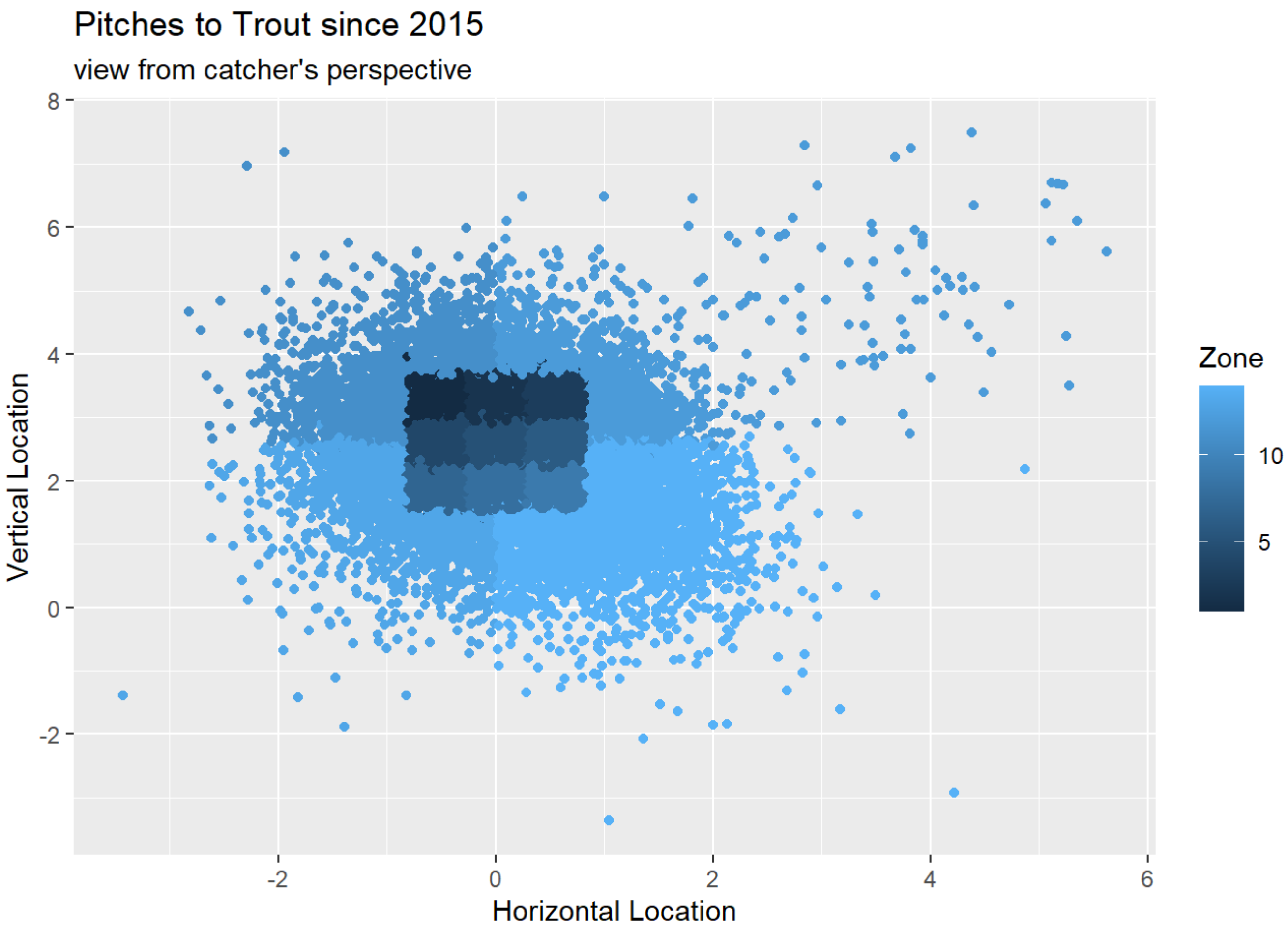
```
Trout %>% filter(is.na(zone))
```

```
## # A tibble: 143 x 7
##    pitch_type description      zone launch_speed plate_x plate_z result
##    <chr>      <chr>           <dbl>        <dbl>   <dbl>   <dbl>  <dbl>
##  1 <NA>       foul_tip           NA            0      NA      NA      0
##  2 <NA>       ball               NA            0      NA      NA      1
##  3 <NA>       foul               NA            0      NA      NA      0
##  4 <NA>       called_strike      NA            0      NA      NA      0
##  5 <NA>       swinging_strike    NA            0      NA      NA      0
##  6 <NA>       ball               NA            0      NA      NA      1
##  7 <NA>       ball               NA            0      NA      NA      1
##  8 <NA>       ball               NA            0      NA      NA      1
##  9 <NA>       ball               NA            0      NA      NA      1
## 10 <NA>       hit_into_play      NA            0      NA      NA      0
## # ... with 133 more rows
```

Let's first see what the zones look like. This scatterplot is shows the location where each pitch was thrown, and is colored by zone.

```
Trout %>% ggplot() +
          geom_point(mapping = aes(x = plate_x, y = plate_z, color = zone)) +
          labs(title = "Pitches to Trout since 2015", subtitle = "view from catcher's perspectiv
e", color = "Zone", x = "Horizontal Location", y = "Vertical Location")
```

```
## Warning: Removed 143 rows containing missing values (geom_point).
```



There is a clear rectangle of the strike zone defined by the 9 blue colors in the middle of this plot. Statcast defines zone 1 as the dark blue upper left section and indexes left to right. The 4 shades of light blue that are outside of the strike zone are 11, 12, 13, and 14.

This means there’s a fundamental difference between zones 1-9 and 11-14; Trout cannot take pitches in zone 1-9 for a ball, so the only way to get a good outcome is to hit the pitch hard. I will compare the individual averages of zone 1-9 with the total average of 1-9 and the same with zones 11-14. This code creates a variable with only zone 1-9 pitches and a variable with zone 11-14 pitches

```
Trout_Strikes <- Trout %>% filter(zone != 11, zone != 12, zone != 13, zone != 14)

Trout_Balls <- Trout %>% filter(zone == 11 | zone == 12 | zone == 13 | zone == 14)
```

Let’s compare the zone 1-9 averages and see how they differ from the total 1-9 result average. The 1-9 total average is:

```
mean(Trout_Strikes$result)
```

```
## [1] 0.191344
```

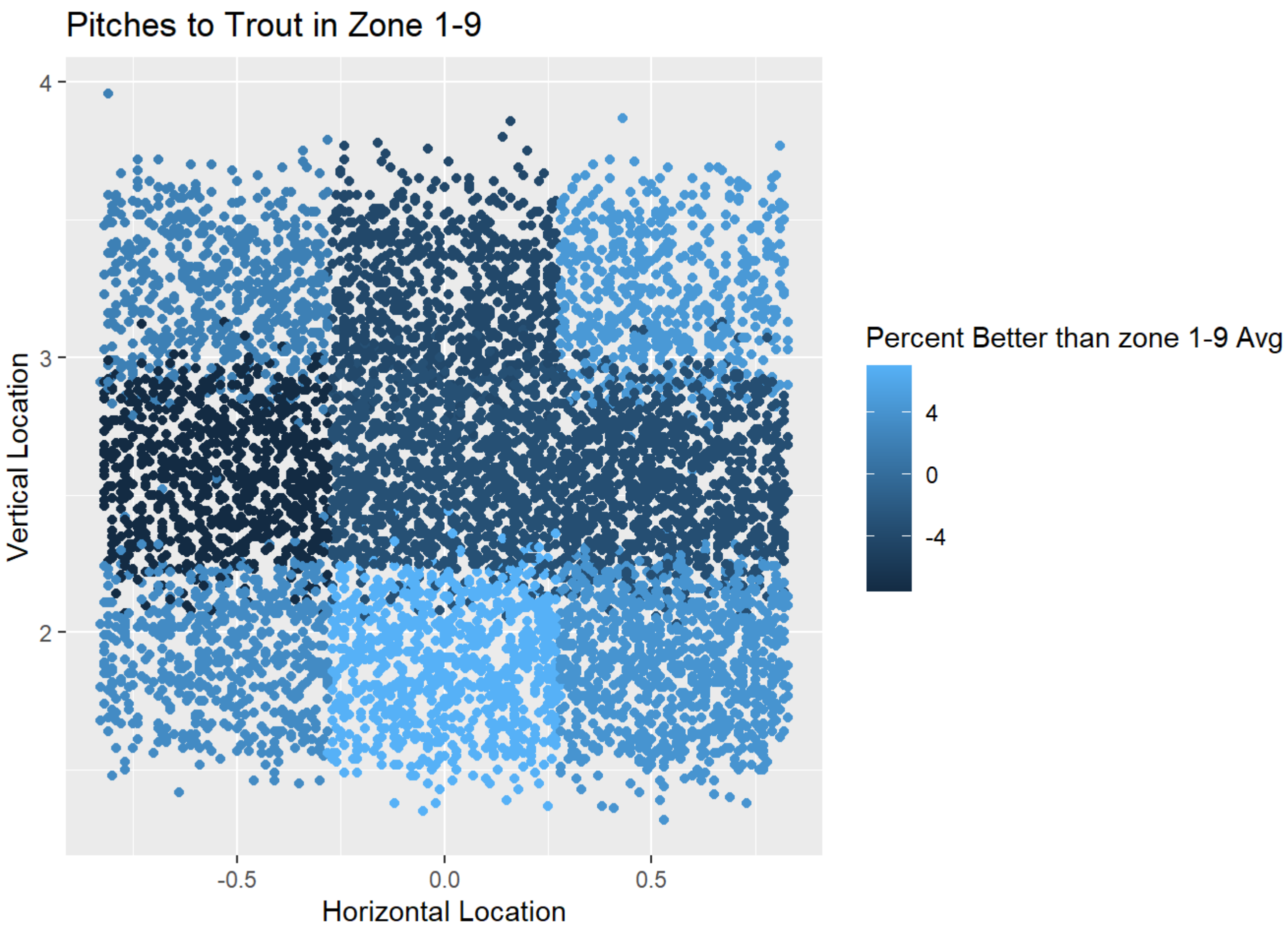
And this code gives individual result averages, as well as what percent better or worse his performance was compared to his performance across all zones.

```
OneNine_Avg <- Trout_Strikes %>% group_by(zone) %>% summarize(mean = mean(result),
                                               vs_avg = (mean - mean(Trout_Strikes$result)) * 10
0,
                                               count = n()
                                               )
OneNine_Avg
```

```
## # A tibble: 9 x 4
##    zone  mean vs_avg count
##   <dbl> <dbl>  <dbl> <int>
## 1     1 0.212   2.04   595
## 2     2 0.151  -4.01   734
## 3     3 0.237   4.61   577
## 4     4 0.116  -7.49   799
## 5     5 0.160  -3.14   988
## 6     6 0.158  -3.31  1005
## 7     7 0.224   3.22   568
## 8     8 0.262   7.04   829
## 9     9 0.233   4.12   929
```

Let’s join these results with the strikes data set to graph and visualize

```
Trout_Strikes %>% left_join(OneNine_Avg, by = "zone") %>%
  ggplot() +
  geom_point(mapping = aes(x = plate_x, y = plate_z, color = vs_avg)) +
  labs(title = "Pitches to Trout in Zone 1-9",  color = "Percent Better than zone 1-9 Avg", x =
"Horizontal Location", y = "Vertical Location")
```



The darker sections are the zones Trout hits worse compared to his average in zone 1-9. It appears that Trout is comparatively worse at hitting the middle-inside pitch, zone 4, and best at the low-middle pitch, zone 8.

Let's repeat the same process with pitches outside the zone.

```
mean(Trout_Balls$result)
```

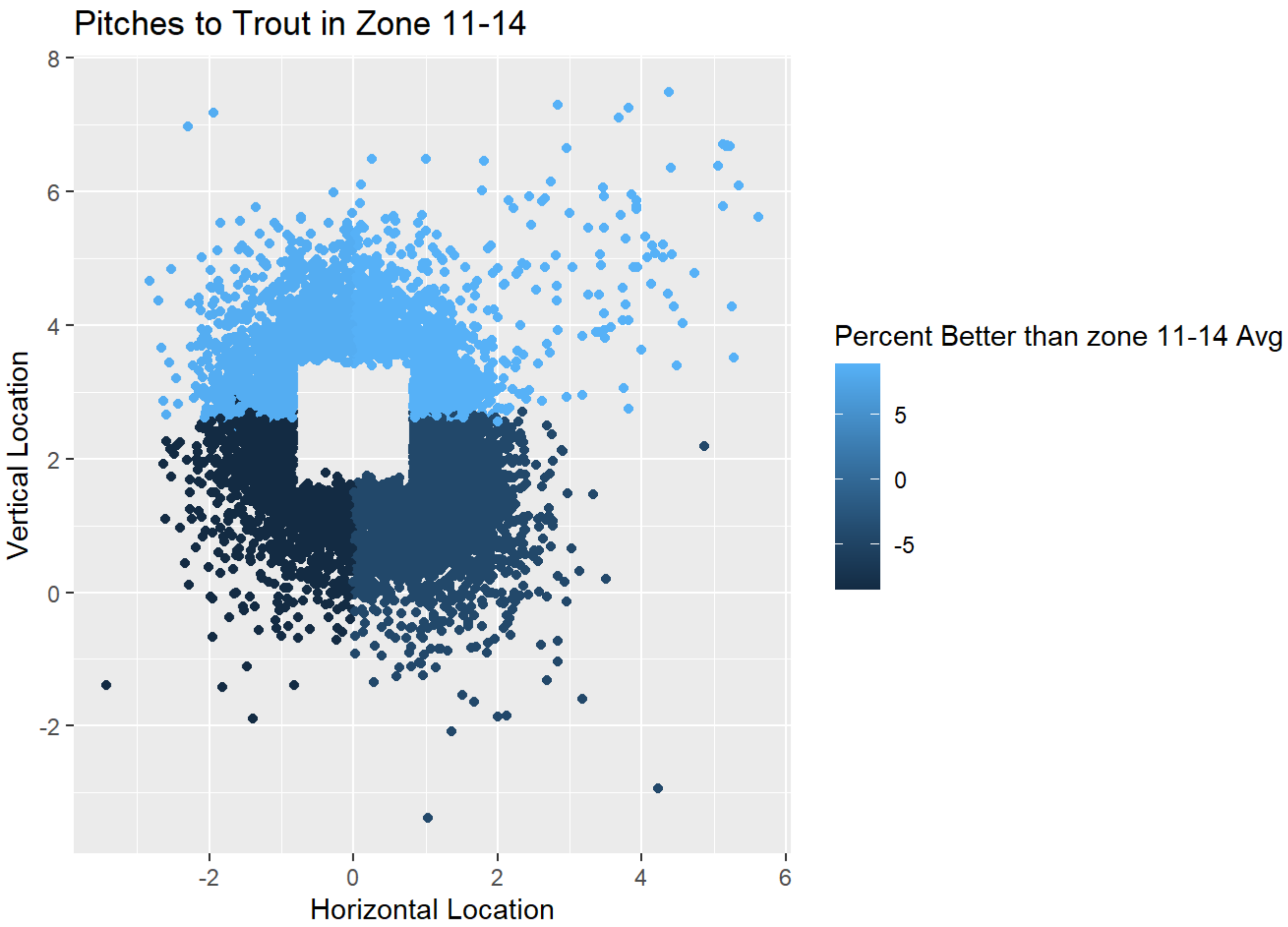
```
## [1] 0.7669852
```

```
ElevFourt_Avg <- Trout_Balls %>% group_by(zone) %>% summarize(mean = mean(result),
                                              vs_avg = (mean - mean(Trout_Balls$result)) * 100,
                                              count = n()
                                              )

ElevFourt_Avg
```

```
## # A tibble: 4 x 4
##    zone  mean vs_avg count
##   <dbl> <dbl>  <dbl> <int>
## 1    11 0.851   8.43  1520
## 2    12 0.856   8.90  1472
## 3    13 0.683  -8.43  1434
## 4    14 0.724  -4.31  3213
```

```
Trout_Balls %>% left_join(ElevFourt_Avg, by = "zone") %>%
  ggplot() +
  geom_point(mapping = aes(x = plate_x, y = plate_z, color = vs_avg)) +
  labs(title = "Pitches to Trout in Zone 11-14",  color = "Percent Better than zone 11-14 Avg",
 x = "Horizontal Location", y = "Vertical Location")
```



While Trout performs quite well on pitches outside of the zone, he performs much worse comparatively on pitches that are low. Zone 13 and 14 are the low-inside and low-outside quadrants.

What could I tell a pitcher based on this information? The most obvious piece of advice would be to pitch into the strike zone, zones 1-9. The pitcher would, of course, laugh at me - strikes being good is not a new idea. However, if the pitcher misses the strike zone it is better to miss low than it is to miss high. When throwing strikes, locating that pitch in zone 4 and zone 2 seem like the best zones for a pitcher to target and zone 8 would be a place to avoid.

Question 2: What pitch types does Mike Trout hit the best/worst?

These are all the different types of pitches Mike Trout saw.

```
Trout %>% count(pitch_type) %>% arrange(desc(n))
```

```
## # A tibble: 15 x 2
##    pitch_type     n
##    <chr>      <int>
##  1 FF          5908
##  2 SL          2337
##  3 SI          1513
##  4 FT          1265
##  5 CH          1064
##  6 CU          1010
##  7 FC           981
##  8 KC           301
##  9 FS           202
## 10 IN            87
## 11 KN            77
## 12 <NA>          53
## 13 EP             6
## 14 CS             1
## 15 PO             1
```

FF refers to a four seam fastball, SL refers to slider, SI refers to sinker, etc. I could go through all of these, but I don’t feel like naming them does much for comprehension at this point. Let’s look at the same mean as we did for zones.

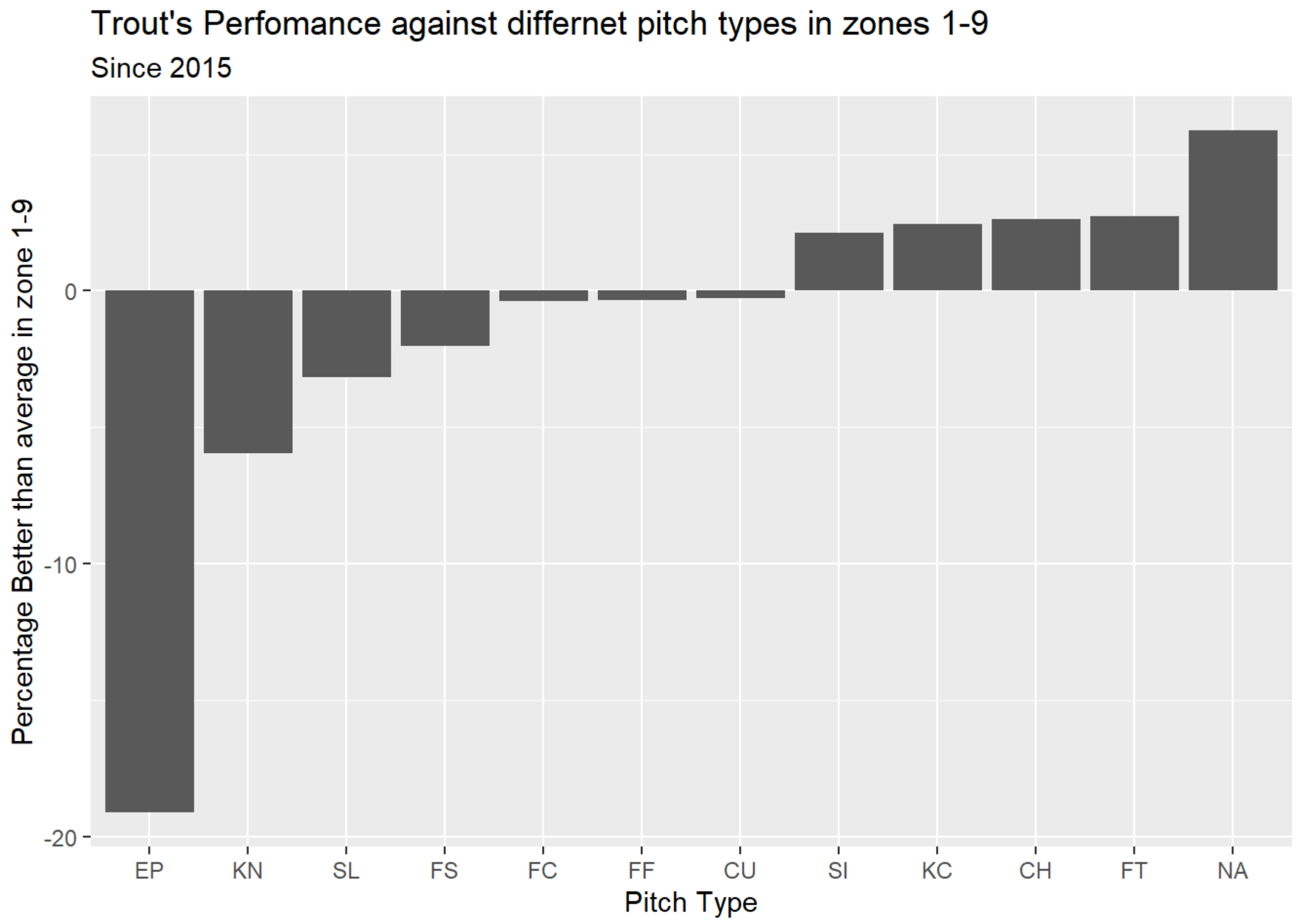
Similar to pitch zones, I will only compare pitches within zone 1-9 with each other and zone 11-14 with each other.

```
Trout_Strikes %>% group_by(pitch_type) %>% summarize(mean = mean(result),
                                                     vs_avg = (mean - mean(Trout_Strikes$resul
t)) * 100,
                                                     count = n()) %>%
  arrange(desc(count))
```

```
## # A tibble: 12 x 4
##    pitch_type  mean  vs_avg count
##    <chr>      <dbl>   <dbl> <int>
##  1 FF         0.188  -0.356  3094
##  2 SL         0.160  -3.16    958
##  3 SI         0.213   2.14    785
##  4 FT         0.218   2.71    650
##  5 FC         0.187  -0.398   475
##  6 CU         0.189  -0.262   408
##  7 CH         0.218   2.64    395
##  8 KC         0.216   2.45    139
##  9 FS         0.171  -2.03     76
## 10 KN         0.132  -5.98     38
## 11 <NA>       0.25    5.87      4
## 12 EP         0     -19.1       2
```

```
Trout_Strikes %>% group_by(pitch_type) %>% summarize(mean = mean(result),
                                                     vs_avg = (mean - mean(Trout_Strikes$resul
t)) * 100,
                                                     count = n()) %>%
  ggplot() +
  geom_histogram(mapping = aes(x = fct_reorder(pitch_type, vs_avg), y = vs_avg), stat = "identit
y") +
  labs(title = "Trout's Perfomance against differnet pitch types in zones 1-9", subtitle = "Sinc
e 2015", x = "Pitch Type", y = "Percentage Better than average in zone 1-9")
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

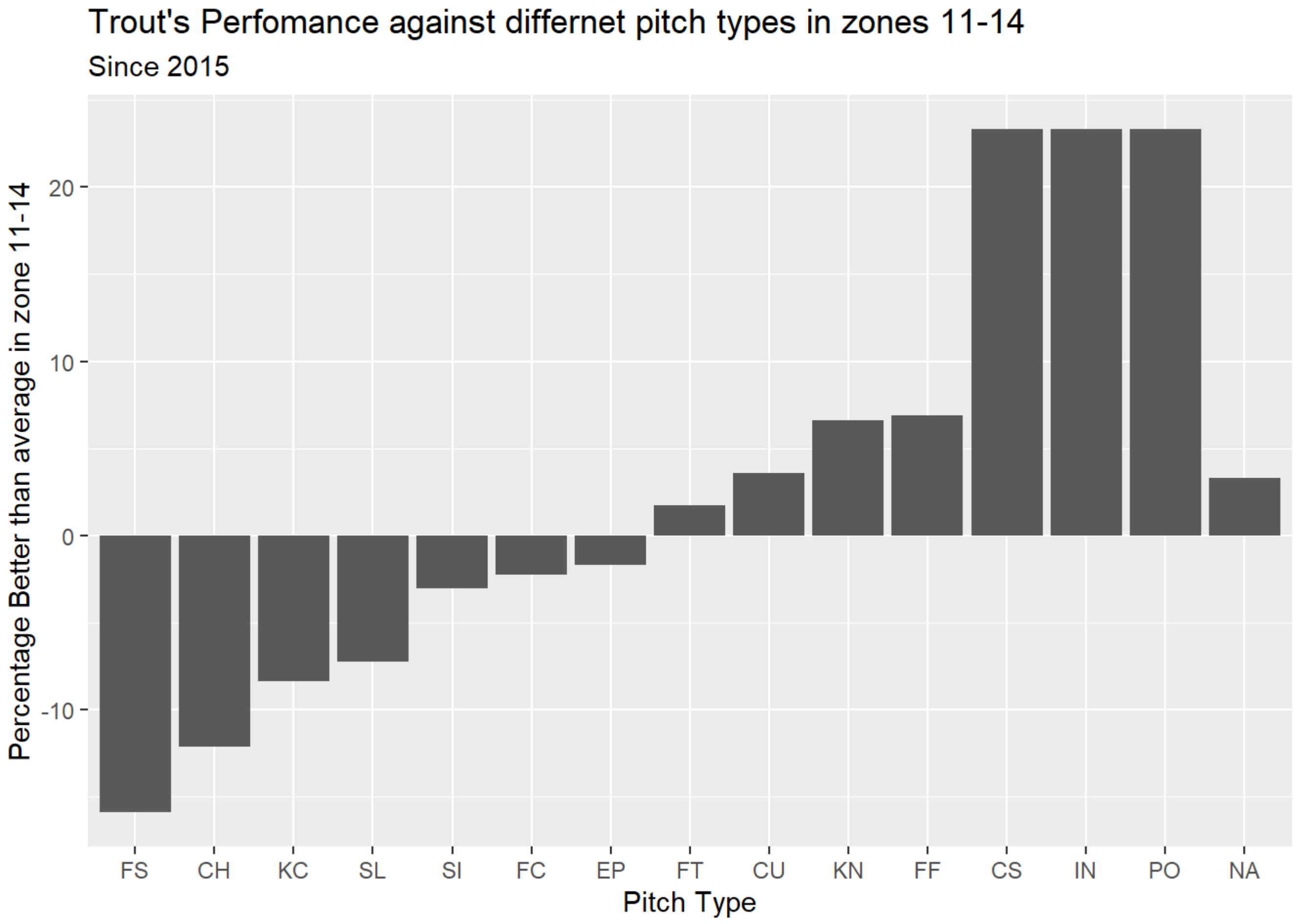


It seems like Trout performs awfully against EP, the eephus, and KN, the knuckleball, pitches he only saw 2 and 38 times respectively. We will ignore those. I think the most interesting takeaways are that he performs worse than average against sliders (SL) and splitters (FS) inside the zone, and good against sinkers (SI), changeups (CU), and two-seam fastballs (FT) inside the zone.

Let’s peform the same visualization for pitches outside of the zone.

```
Trout_Balls %>% group_by(pitch_type) %>% summarize(mean = mean(result),
                                                     vs_avg = (mean - mean(Trout_Balls$result))
* 100,
                                                     count = n()) %>%
  ggplot() +
  geom_histogram(mapping = aes(x = fct_reorder(pitch_type, vs_avg), y = vs_avg), stat = "identit
y") +
  labs(title = "Trout's Perfomance against differnet pitch types in zones 11-14", subtitle = "Si
nce 2015", x = "Pitch Type", y = "Percentage Better than average in zone 11-14")
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



The categories IN and PO both refer to instances where the pitcher intentionally throws a ball, and makes that intention clear. The reasons for that choice aren't important here - I mention it to say that they are not pitch types we are interested in.

I'm focused on Trout being better at dealing with four-seam fastballs outside the strike zone than he is at dealing with Splitters (FS), Changeups (CH), Knuckle-Curves (KC), sliders (SL), and Sinkers (SI). What all of the pitches he does worse with have in common is some kind of spin-induced movement. For example, sliders (from a right-handed pitcher) move down and away from him and Knuckle-curveballs drop directly downward. I speculate that he is worse with these moving pitches because they initially appear they are headed to the strike zone, and then move out of that path. A fastball travels straighter so that initial perception would be more accurate. This could also be why it's better to miss the strike zone low; these pitches all move downward (or "sink") in some way, meaning that if they moved out of the path towards the strike zone, they'd move to end up below it.

Question 3: Based on pitch location and type, how well could you expect Mike Trout to hit?

I'm going to narrow down the pitch types we look at. Because Trout has seen fewer of certain pitch types, he hasn't seen them a significant amount of times in specific zones. I'm going to look at the three most commonly thrown pitches - FF, SL, and SI. I think these are good choices for analysis because he performs differently against each of them - compared to all pitches thrown, he is about average against 4-seam fastballs (FF), below average against sliders (SL), and above average against sinkers (SI).

This code filters the dataset to only include pitches of those pitch_types

```
Trout_MostPitches <- Trout %>% filter(pitch_type == "FF" | pitch_type == "SL" | pitch_type == "S
I")
```

This generates a model that includes success rate predictions based on the pitch_type and zone.

```
Result_Model <- glm(result ~ pitch_type*zone, family = "binomial", data = Trout_MostPitches)
```

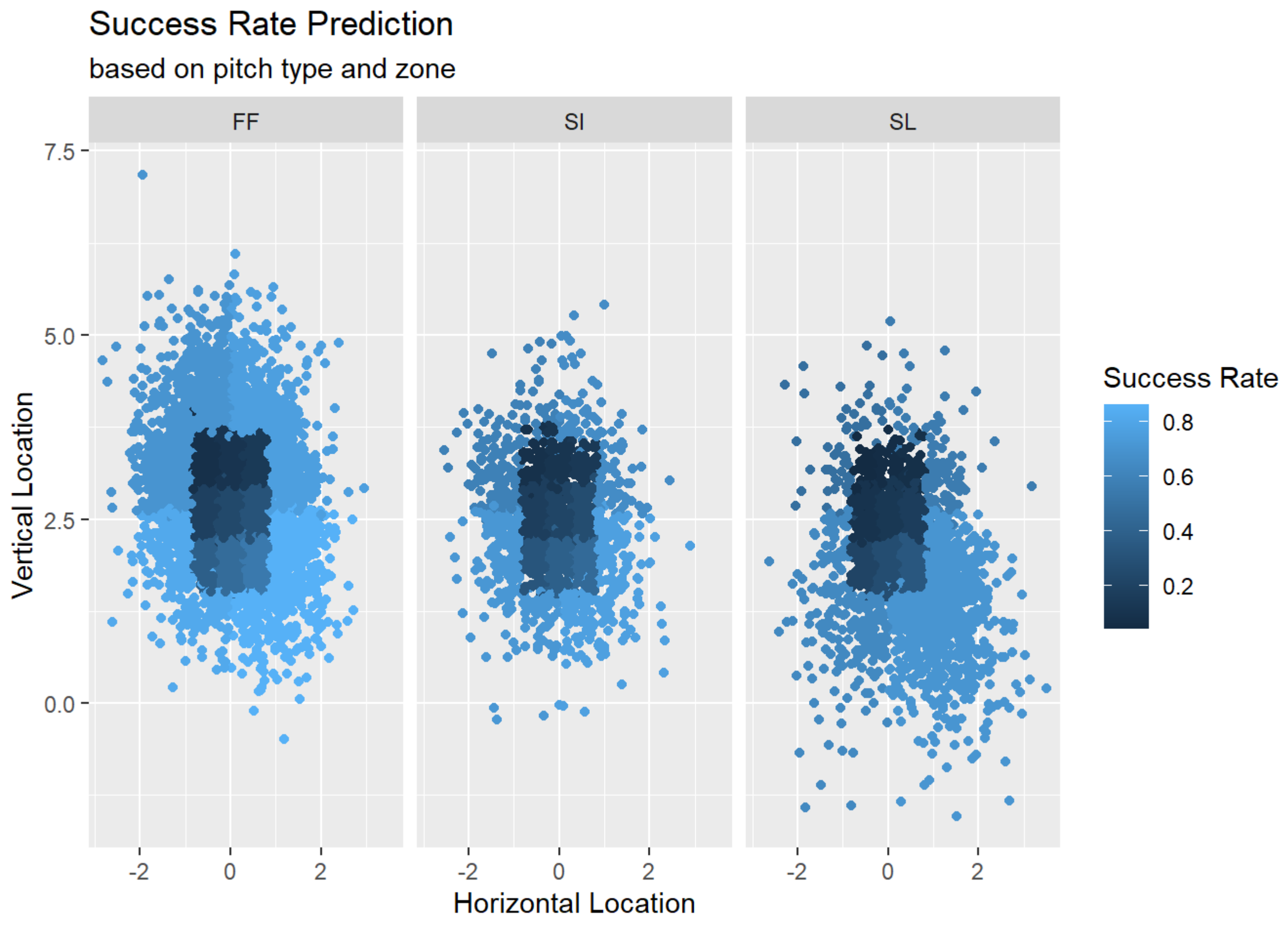
This code adds those predictions to the new object Trout_Model

```
Trout_Model <- Trout_MostPitches %>% add_predictions(Result_Model, var = "p_pred", type = "respo
nse")
```

This shows the success rate prediction for each different pitch

```
Trout_Model %>% ggplot() +
  geom_point(mapping = aes(x = plate_x, y = plate_z, color = p_pred)) +
  facet_wrap(~pitch_type) +
  labs(title = "Success Rate Prediction", subtitle = "based on pitch type and zone", color = "Su
ccess Rate", x = "Horizontal Location", y = "Vertical Location")
```

```
## Warning: Removed 62 rows containing missing values (geom_point).
```



It looks like high strikes are most effective across the board and there also doesn't seem to be much variance based on pitch type. There are some useful takeaways. Four-seam fastballs are most effective in zones 1-3. I think this is an interesting insight because when we looked only at zone, zone 4 his worst. The strategy of throwing high four-seem fastballs seems to be reflected by the pitch locations; there are far more FF pitches above the zone compared to sinkers or sliders. However, this reflection doesn't seem to happen with sliders. This model predicts that Trout will hit sliders in zone 9 the best of all in-strike zone zones, but there appears to be a lot of pitches thrown around that spot. Perhaps there isn't enough observations of high sliders to have a lot of confidence in the models prediction that high and inside sliders (zones 1-4) are the most effective locations. Sinkers look to be about the same as four-seam fastballs.

Final Discussion: Takeaways, Ethics, and Issues

Broadly, it looks like pitchers should throw inside to Mike Trout. More specifically, it seems like high-fastballs and sliders are the most effective pitches. It would be a bad idea to throw him sinkers, two-seamers, and changeups.

I think it's hard to find a substantial ethical concern. In terms of data collection, the games that this data is from are publicly broadcasted, and, for example, the launch speed of the ball off Trout's bat is not sensitive or private information. In terms of results, it's hard to find a concern that doesn't seem overly confident in my findings. Let's pretend that these findings are 100% accurate, and that if pitchers used this information effectively, Trout would be a less successful hitter. What might happen as a result of Trout being less successful? He wouldn't be paid less - he has already signed a multi-million dollar contract the guarantees him money. His team might perform worse as a whole. Even if that does happen, I don't think you can argue that performing this analysis was unethical, because it’s a component of baseball competition. Hitters like Trout use similar reports on the pitcher they face to try to gain an advantage. I don’t see any other far-reaching affects of this study that could be considered ethically significant.

The main issue with this analysis is my definition of effective. A pitch hit greater than 95 miles an hour is far more valuable than a pitch taken for a ball, but my binary metric equates them. I also count taking a pitch that is a strike as bad, while there are many reasons why a batter may choose not to swing. They could have been guessing a certain pitch was coming and happy to let a pitch pass that they wouldn’t have hit well anyways.

I think a larger and more accurate data set would be helpful. There were discernibly less slider thrown in the upper part of the zone compared to four-seam fastballs (as well as twice as many four-seamers regardless of location). More pitches in the dataset would also mean we could look at more than 3 pitch types in the third question.

I also think it would be helpful to look at more than just Mike Trout. I think that would help validate the accuracy of my metric as a marker for offensive value. If the leaderboard for my stat looks similar to the leaderboard for other commonly used offensive statistics, I think it’d be fairly accurate. Because I don’t have that information, it’s hard to answer the question of what makes Trout great - we don’t know if how his averages compare to league-wide averages.

Thank you for reading!