Stealing Discovery Analysis

Dusty Turner

June 13, 2019

Data

1. The data we are analyzing below come from the following website: Baseball Savant
2. I learned by trial and error that you can only download 40,000 rows of data at a time. You'll have to add it together to do more stuff.

Read in Data and Set Options

```
library(tidyverse)
library(stringr)
options(tibble.print_min = 100)
data = read_csv("savant_data_stealing.csv",guess_max = 10000) %>% janitor::clean_names()
# data = read_csv("savant_data_stealing_everypitch.csv",guess_max = 10000) %>% janitor::clean_names()
```

The data has the following columns available:

```
names(data)
```

```
##  [1] "pitch_type"                "game_date"
##  [3] "release_speed"             "release_pos_x"
##  [5] "release_pos_z"             "player_name"
##  [7] "batter"                    "pitcher"
##  [9] "events"                    "description"
## [11] "spin_dir"                  "spin_rate_deprecated"
## [13] "break_angle_deprecated"    "break_length_deprecated"
## [15] "zone"                      "des"
## [17] "game_type"                 "stand"
## [19] "p_throws"                  "home_team"
## [21] "away_team"                 "type"
## [23] "hit_location"              "bb_type"
## [25] "balls"                     "strikes"
## [27] "game_year"                 "pfx_x"
## [29] "pfx_z"                     "plate_x"
## [31] "plate_z"                   "on_3b"
## [33] "on_2b"                     "on_1b"
## [35] "outs_when_up"              "inning"
## [37] "inning_topbot"             "hc_x"
## [39] "hc_y"                      "tfs_deprecated"
## [41] "tfs_zulu_deprecated"       "fielder_2"
## [43] "umpire"                    "sv_id"
## [45] "vx0"                       "vy0"
## [47] "vz0"                       "ax"
## [49] "ay"                        "az"
## [51] "sz_top"                    "sz_bot"
## [53] "hit_distance_sc"           "launch_speed"
## [55] "launch_angle"              "effective_speed"
## [57] "release_spin_rate"         "release_extension"
```

```
## [59] "game_pk"                          "pitcher_1"
## [61] "fielder_2_1"                      "fielder_3"
## [63] "fielder_4"                        "fielder_5"
## [65] "fielder_6"                        "fielder_7"
## [67] "fielder_8"                        "fielder_9"
## [69] "release_pos_y"                    "estimated_ba_using_speedangle"
## [71] "estimated_woba_using_speedangle" "woba_value"
## [73] "woba_denom"                       "babip_value"
## [75] "iso_value"                        "launch_speed_angle"
## [77] "at_bat_number"                    "pitch_number"
## [79] "pitch_name"                       "home_score"
## [81] "away_score"                       "bat_score"
## [83] "fld_score"                        "post_away_score"
## [85] "post_home_score"                  "post_bat_score"
## [87] "post_fld_score"                   "if_fielding_alignment"
## [89] "of_fielding_alignment"
```

The dates contain the following range:

```
data$game_date %>% range()
```

```
## [1] "2019-05-19" "2019-06-11"
```

Based on the way I requested the data, the data only contains plate appearances where there was at least one runner on base

Determining Who Stole Second

My apologies for not commenting this well. I recomend running this line by line to determing what is happening.

```
##
stolesecond =
data %>%
  mutate(pitchID = row_number()) %>%
  # select(pitchID, events,des,pitch_type,release_speed,batter,at_bat_number,on_1b,on_2b,on_3b) %>%
  select(pitchID, batter, pitcher,balls, strikes,at_bat_number,on_1b,on_2b,on_3b) %>%
  mutate_at(c("on_1b","on_2b","on_3b"), as.numeric) %>%
  mutate(on_1b = na_if(on_1b,"null"), on_2b = na_if(on_2b,"null"), on_3b = na_if(on_3b,"null")) %>%
  group_by(batter,pitcher,at_bat_number) %>%
  arrange(batter) %>%
  filter(sum(is.na(on_2b))>1&&sum(!is.na(on_2b))>1) %>%
  group_by(batter,pitcher,at_bat_number,on_2b) %>%
  mutate(groupID = row_number())  %>%
  ungroup() %>%
  filter(!is.na(on_1b)||!is.na(on_2b)) %>%
  filter(!is.na(on_2b)) %>%
  group_by(batter,pitcher,at_bat_number,on_2b) %>%
  filter(groupID == max(groupID)) %>% ## this is the pitch he stole from 1st to second
  ungroup() %>%
  select(pitchID,batter) %>%
  rename(stole2nd = batter)
```

Look at it:

```
stolesecond %>% head
```

```
## # A tibble: 6 x 2
```

```
##   pitchID stole2nd
##     <int>    <dbl>
## 1   37017   408234
## 2   21640   425772
## 3   11672   425783
## 4   10559   429665
## 5   30643   430945
## 6   20034   431145
```

```
stolesecond %>% count(stole2nd,sort = TRUE) %>% top_n(10)
```

```
## # A tibble: 14 x 2
##    stole2nd     n
##       <dbl> <int>
##  1   467793     5
##  2   547379     4
##  3   605141     4
##  4   452254     3
##  5   452678     3
##  6   502054     3
##  7   547180     3
##  8   570560     3
##  9   571718     3
## 10   571745     3
## 11   595284     3
## 12   596129     3
## 13   605233     3
## 14   621020     3
```

Now that we've discover who has stole second, lets add that data back into the main dataset.

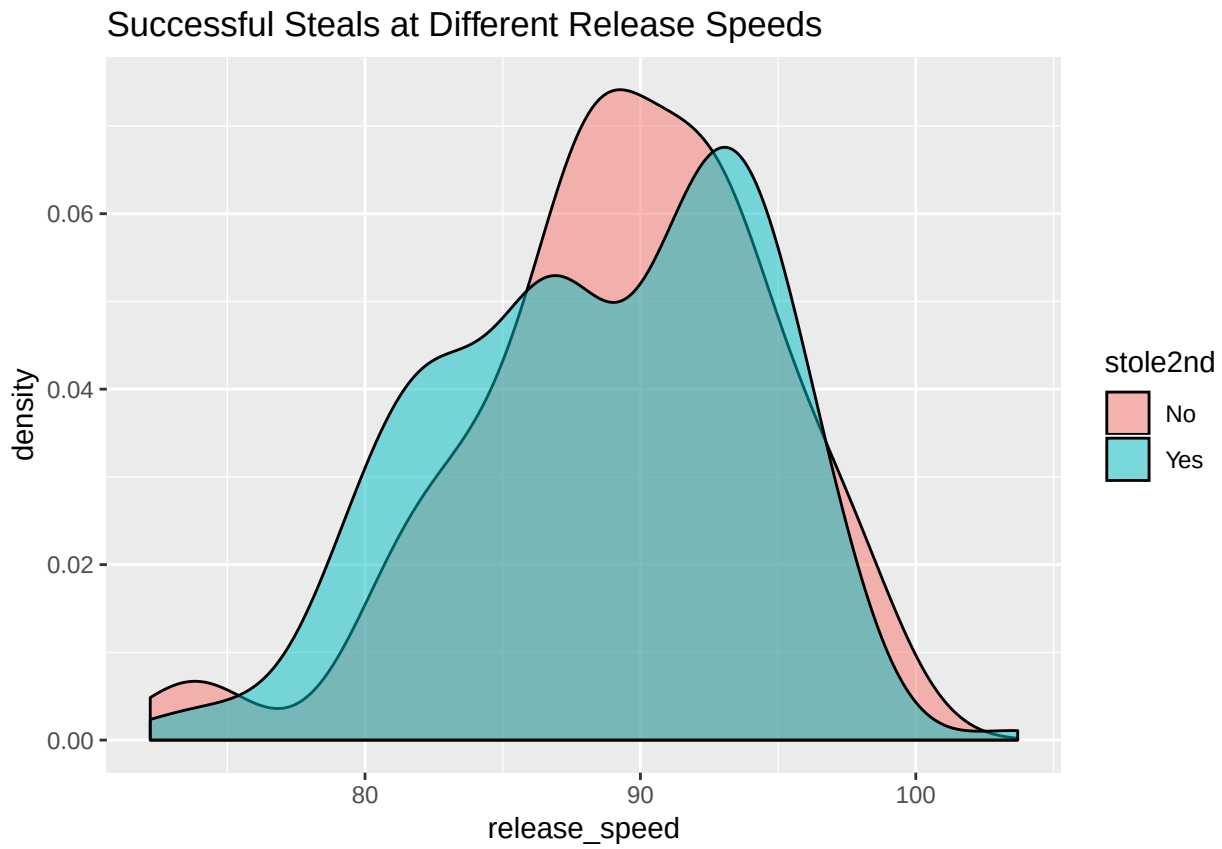
```
secondseal =
data %>%
  # select(game_date,batter, des,pitcher,balls, strikes,outs_when_up,at_bat_number,on_1b,on_2b,on_3b, r
  mutate(pitchID = row_number(), des = na_if(des,"null")) %>%
  left_join(stolesecond) %>%
  mutate(stole2nd = ifelse(is.na(stole2nd), "No","Yes")) %>%
  mutate(stole2ndcaught = ifelse(str_detect(des,"caught")&str_detect(des,"2nd base"), "Yes", "No")) %>%
  mutate(stealattempt = ifelse(stole2nd=="Yes"|stole2ndcaught=="Yes","Yes","No")) %>%
  mutate(release_speed = as.numeric(release_speed)) %>% filter(stealattempt == "Yes")
  # select(pitchID,des, stole2nd,stole2ndcaught,stealattempt)
```

Look at data:

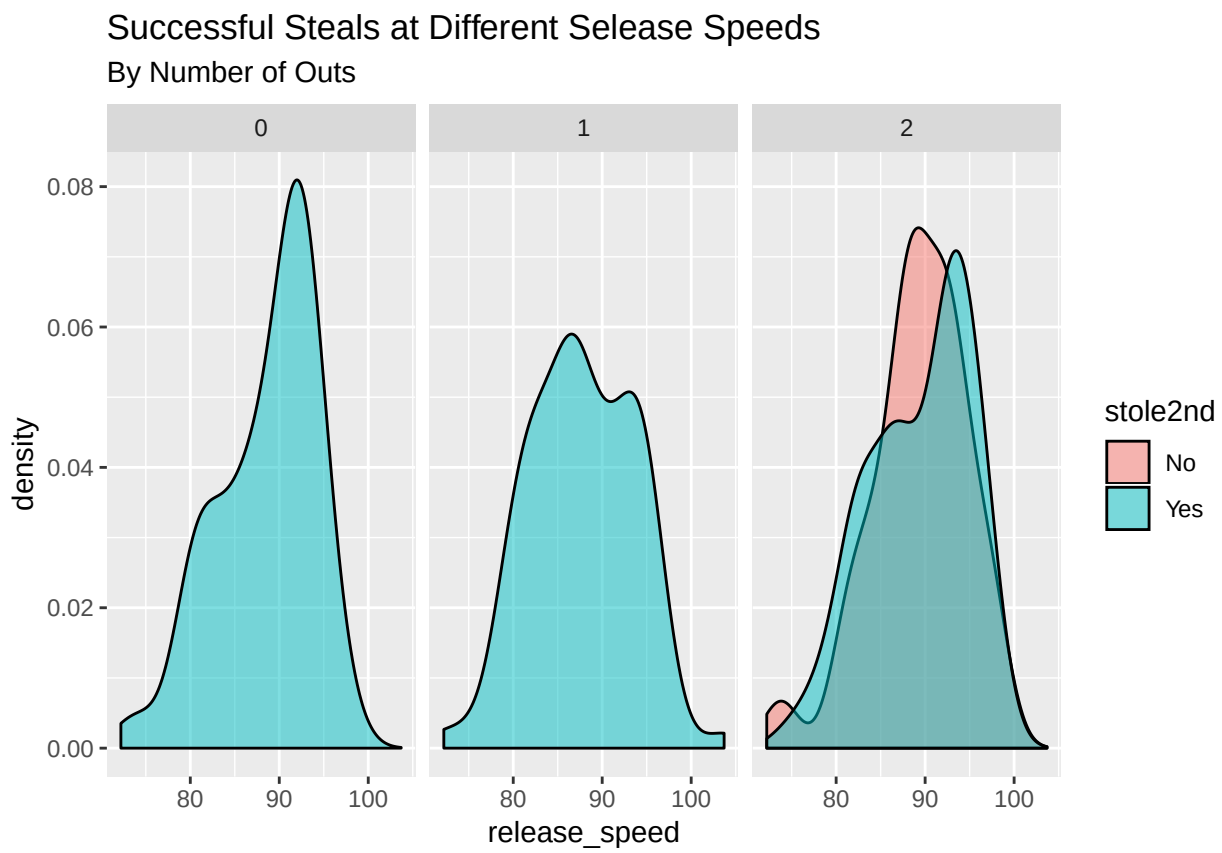
```
secondseal %>% head
```

Visualizations

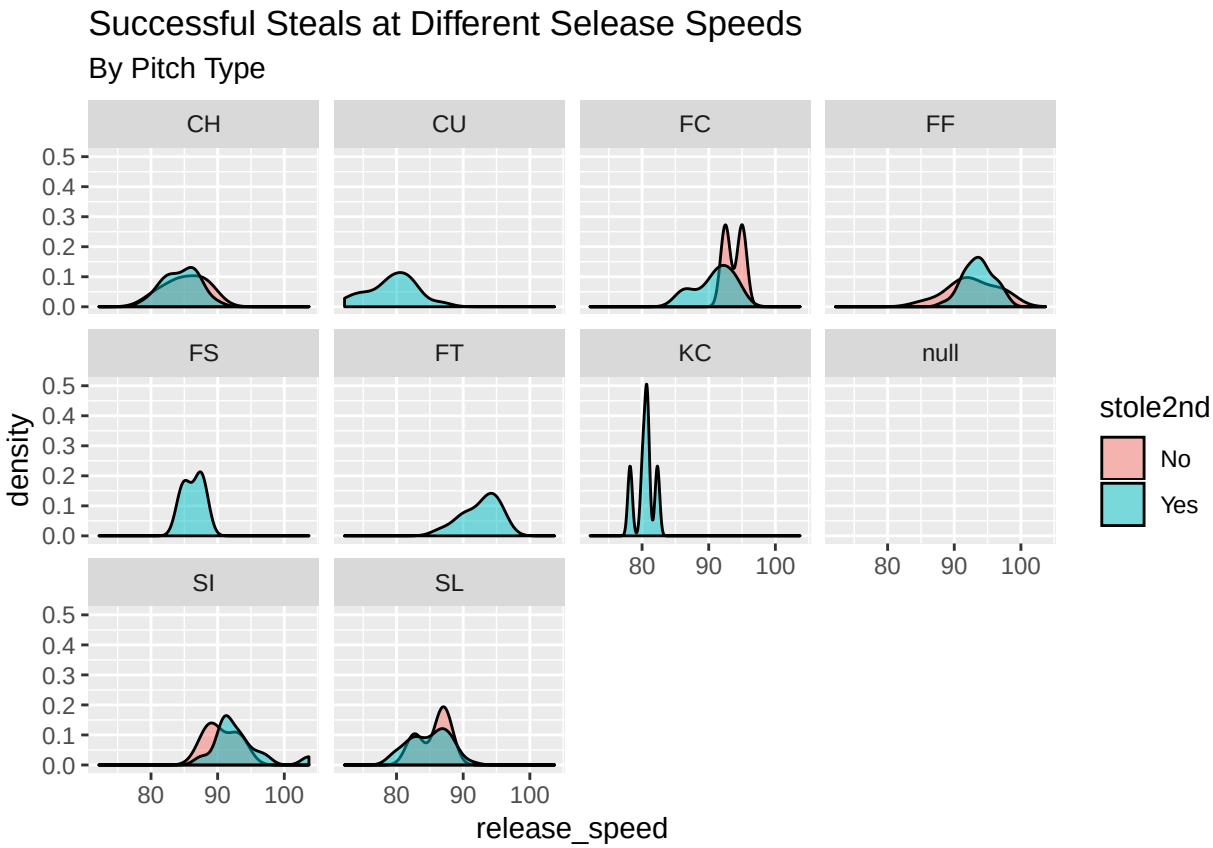
```
secondseal %>%
  ggplot(aes(x=release_speed, fill = stole2nd)) + geom_density(, alpha = .5) +
  labs(title = "Successful Steals at Different Release Speeds")
```




```
secondseal %>%
  # select(pitch_type,release_speed,stole2nd) %>%
  select(release_speed,stole2nd, outs_when_up) %>%
  gather(key, other,-release_speed,-stole2nd) %>%
  ggplot(aes(x=release_speed, fill = stole2nd)) + geom_density(, alpha = .5) +
  facet_wrap(~other) +
  labs(title = "Successful Steals at Different Selease Speeds", subtitle = "By Number of Outs")
```




```
secondseal %>%
  select(pitch_type,release_speed,stole2nd) %>%
  # select(release_speed,stole2nd, outs_when_up) %>%
  gather(key, other,-release_speed,-stole2nd) %>%
  ggplot(aes(x=release_speed, fill = stole2nd)) + geom_density(, alpha = .5) +
  facet_wrap(~other) +
  labs(title = "Successful Steals at Different Selease Speeds", subtitle = "By Pitch Type")
```



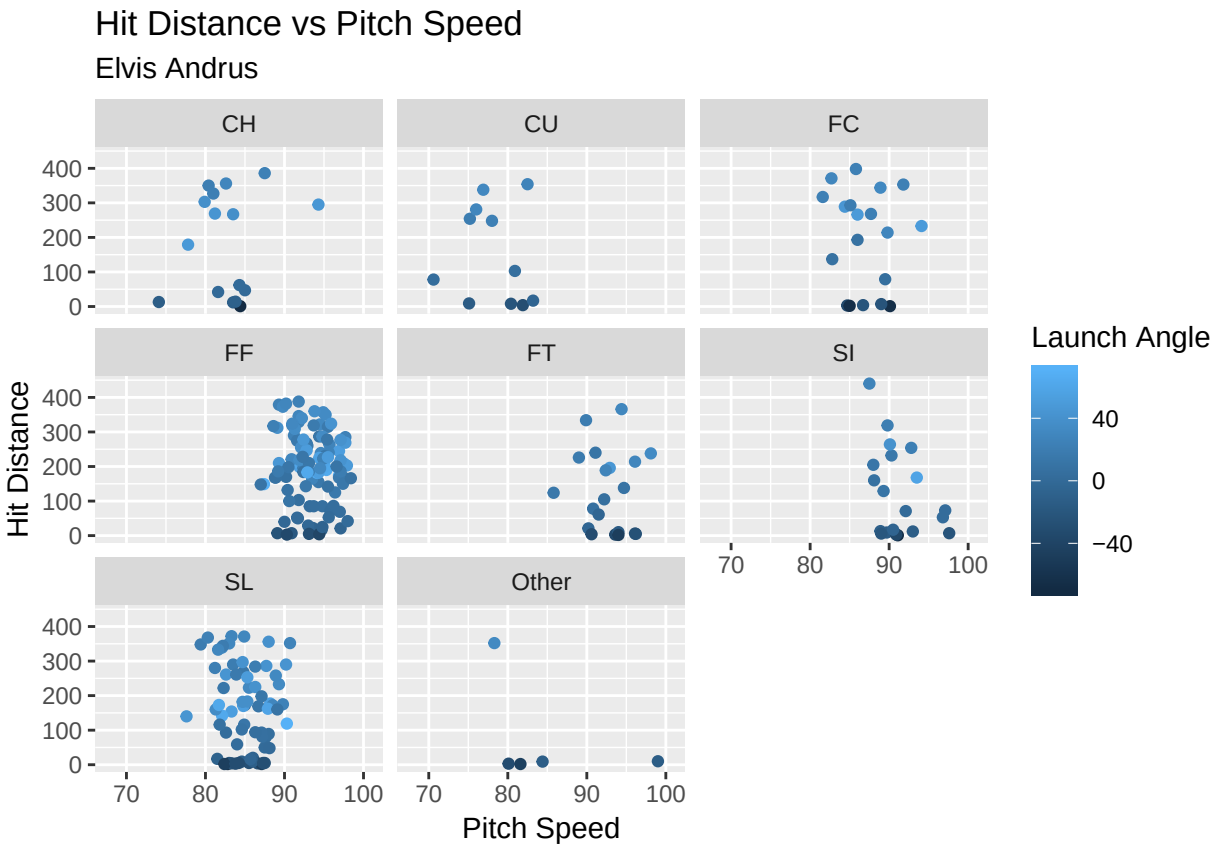
Stealing Third Attempts

Not currently correct

A Look at Other Batting Stuff

```
texas = read_csv("texas_batters.csv") %>% janitor::clean_names() %>%
  mutate(release_speed = as.double(release_speed),
         launch_speed = as.double(launch_speed),
         launch_angle = as.double(launch_angle),
         # launch_angle = as.double(launch_angle),
         hit_distance_sc = as.double(hit_distance_sc),
         hc_x = as.double(hc_x),
         hc_y = as.double(hc_y))


texas %>%
  select(player_name,release_speed,launch_speed,pitch_type,launch_angle,hit_distance_sc) %>%
  filter(player_name=="Elvis Andrus") %>%
  mutate(pitch_type = fct_lump(pitch_type)) %>%
  ggplot(aes(x=release_speed, y=hit_distance_sc)) +
  geom_point(aes(color = launch_angle)) +
  facet_wrap(~pitch_type) +
  labs(title = "Hit Distance vs Pitch Speed", x="Pitch Speed", y = "Hit Distance",
       color = "Launch Angle", subtitle = "Elvis Andrus")
```




```
texas %>%
  select(player_name,release_speed,launch_speed,pitch_type,
         launch_angle,hit_distance_sc, hc_x,hc_y,stand,of_fielding_alignment) %>%
  # filter(player_name=="Elvis Andrus") %>%
```

```
  mutate(of_fielding_alignment = fct_lump(of_fielding_alignment)) %>%
  ggplot(aes(x=hc_x, y=hc_y)) +
  geom_point(aes(color = launch_angle)) +
  facet_grid(of_fielding_alignment~stand) +
  labs(title = "Hit Distance vs Pitch Speed", x="Pitch Speed", y = "Hit Distance",
       color = "Launch Angle", subtitle = "Elvis Andrus")
```