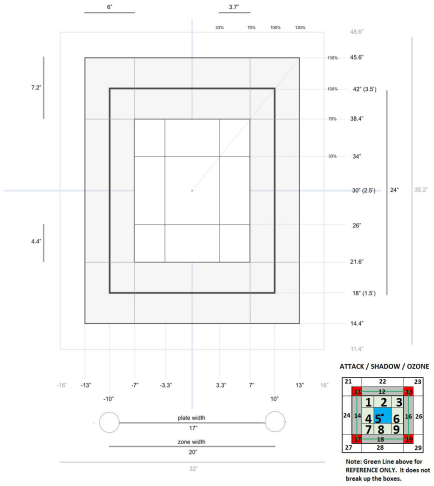
Adam Chapnik
MATH-242
December 12, 2018

"Statcast_2017_2018.csv" Codebook

Major League Baseball has been collecting data on every pitch thrown in its regular season games since 2008. That year, MLB set up a camera system in every ballpark, called Pitch F/X that tracked the velocity of every pitch roughly after release. In 2017, MLB replaced Pitch F/X with Statcast, which was able to track and record the velocity of every pitch immediately at release. In addition to tracking pitch velocities, MLB has been recording pitch types, the outcome of the pitch, the outcome of the play, etc. since 2008. MLB has made this data publicly available via Baseballsavant.com.

For the purposes of this study, I gathered the data for every pitch thrown from 2008 to 2018 in a MLB regular season game via Baseballsavant.com through a combination of scraping and directly downloading the data from the cite. Since only the data on pitch velocity in years 2017 and 2018 was inconsistent with the data from 2008 to 2016, my final data is only for years 2017 and 2018. However, in order to create data on players' histories, I used cumulative data on pitch and play outcomes from 2008 through 2018 for every unique player in the dataset from 2017 to 2018. In addition, the dataframe from MLB was incomplete, so I merged a few other datasets to it: park factors, wOBA and FIP constants, umpire names, and venue names and weather. To this, I added about 900 variables for players' career outcomes. A large fraction of these were for pitch outcomes based on the general area in the strike zone the pitch was thrown (I standardized the strike zone and pitch location based on the strike zone's top and bottom for every batter and divided the strike zone into 25 "zones" based on the instructions here). The zones are numbered, as integers, 1-9, 11-14, 16-19, 21-24, and 26-30, with 30 being absolutely outside of the strike zone, whereas zones 11-29 could be either inside or outside of the strike zone, depending on the umpire's call.

For illustration, here is a diagram of the zone breakdown (zone 30 is outside of the zone)



In all, the data set has 998 unique columns and 1,525,025 unique observations.

The following is a table of the variable names, descriptions, units or format, and ranges or counts. Since many of the columns are minor variations of each other, some variables are "grouped" in one row below and marked with the "(#-#)". The two types of variables that are grouped are "per zone" and "per pitch type" variables. "Per zone" variables have "(1-30)" next to their names below, meaning that there are actually 25 distinct columns within this group for each zone. "Per pitch" variables have "(0-16)" next to their names below, meaning that there are 17 distinct columns within this group for each pitch type.

Variable Name	**Description**	**Units/Format**	**Range or Count**
game_pk	The game identifier.	######	4119 unique values (unique games)
game_date	The date of the game.	yyyy-mm-dd	2017-04-02 to 2018-10-01
home_team	Three character abbreviation for home team.	e.g BOS	30 unique names in total
away_team	Three character abbreviation for away team.	e.g. NYY	(Same as home_team)
venue_name	The name of the ballpark of the home team.	Fenway Park	35 unique names in total
temperature	recorded air temperature in the ballpark at the start of the game, in degrees fahrenheit	continuous, in fahrenheit	27-108
other_weather	basic description of the weather at the start of the game	dome = 0, roof close = 1, sunny = 2, clear = 3, partly cloudy = 4, cloudy = 5,	0-9

		overcast = 6, drizzle = 7, rain = 8, snow = 9	
umpire_hp	name of the umpire behind home plate for the game	First Last	94 unique names in total
pitch_type	type of pitch thrown to the batter if available	0 = four-seam fastball, 6 = two-seam fastball, 3 = cutter, 7 = sinker, 8 = splitter, 1 = curveball, 4 = slider, 12 = knuckleball, 2 = changeup, 5 = knuckle-curveball, 11 = eephus, 14 = screwball, 9 = pitch out, 13 = other, PO = 10	0-14
release_speed	speed of the pitch either at or immediately after release, in miles per hour	continuous, in mph	41-105
release_pos_x	The position of the ball on the x-axis when it is released, measured in feet where 0 is the middle of home plate and negative number go towards the right hand hitter batter's box and positive number go	continuous, in feet	-6.5048 through 6.5676

	towards the left hand hitter batter's box. In other words, right hand pitchers will have negative values while left hand pitchers will have positive values.		
release_pos_z	The height of the ball on the z-axis when it is released, measured in feet where 0 is on the ground.	continuous, in feet	-0.0263 through 8.7559
release_pos_y	The distance of the ball from home plate when it is released, measured in feet.	continuous in feet	28.5787 through 61.4139
pitcher	The pitcher's name.	First Last	905 unique names in total
batter	The batter's name.	First Last	750 unique names in total
events	The result of the play.	0 = no contact, 1 = contact, 2 = homerun, 3 = hit by pitch, 4 = walk, 5 = strikeout, 6 = bunt	0-6
description	The result of the pitch.	0 = no swing, 1 = hit, 2 = whiff, 3 = foul, 4 = unknown strike	0-4
zone	The location of the pitched ball in the adjusted strike zone for the batter (Tango, 2018).	integer: 1-9, 11-14, 16-19, 21-24, 26-30	1-30 (25 in total)

stand	The stance of the batter (left or right).	L = 1, R = 0	0 or 1
p_throws	What arm the pitcher throws with (left or right).	L = 0, R = 1	0 or 1
type	The call on the pitch (ball, strike, or in play).	Ball = 1, Strike = 0, In play = 2	0-2
bb_type	Batted ball type (fly ball, ground ball, line drive, popup, or null).	0 = no contact, 1 = contact	0 or 1
balls	The number of balls in the count before the pitch.	integer	0-3
strikes	The number of strikes in the count before the pitch.	integer	0-2
pfx_x	Horizontal movement of the pitch between the release point and home plate, as compared to a theoretical pitch thrown at the same speed with no spin-induced movement, measured in feet.	continuous in feet	-4.1504 through 2.9312
pfx_z	Vertical movement of the pitch between the release point and home place, as compared to a theoretical pitch thrown at the same speed with no spin-induced movement, measured in feet.	continuous in feet	-4.4712 through 6.3036
vx0	The horizontal velocity along the x-axis of the	continuous in feet per second	-21.8674 through 19.8378

	pitch, in feet per second, measured at release.		
vy0	The horizontal velocity along the y-axis of the pitch, in feet per second, measured at release.	continuous in feet per second	-152.9365 through -57.6575
vz0	The vertical velocity (along the z-axis) of the pitch, in feet per second, measured at release.	continuous in feet per second	-19.7706 through 15.533
ax	The horizontal acceleration along the x-axis of the pitch, in feet per second per second, measured at release.	continuous in feet per second per second	-59.2901 through 32.5002
ay	The horizontal acceleration along the y-axis of the pitch, in feet per second per second, measured at release.	continuous in feet per second per second	-0.5244 through 47.2165
az	The vertical acceleration (along the z-axis) of the pitch, in feet per second per second, measured at release.	continuous in feet per second per second	-77.1714 through 16.9763
effective_speed	The perceived speed of the ball in miles per hour. The actual speed of the pitch is adjusted for how close it is to home plate when it is released.	continuous in miles per hour	36.03 through 194.574

release_spin_rate	The total spin rate of the pitch after it is released, measured in rotations per minute.	continuous in rotations per minute	413 through 3726
release_extension	How far from the rubber the ball is when it’s released, measured in feet.	continuous in feet	-0.916 through 10.4348
at_bat_number	Order of the at bat in the game overall, regardless of which team was hitting.	integer	1-144
pitch_number	Order of the pitch in a given at bat.	integer	1-21
horiz_adj	The adjusted pitch position along the width of the strike zone, using the actual pitch position in terms of the width of the plate, in percent.	continuous, in "percent"	-100 through 100
vert_adj	The adjusted pitch position along the height of the strike zone, using the actual pitch position in terms of the top and bottom of the batter's strike zone, in percent.	continuous, in "percent"	-100 through 100
birth_year	The year the batter was born.	year, 19##	1973-1998
pit_birth_year	The year the pitcher was born.	year, 19##	1973-1997

wind_dir	The average general direction the wind is blowing over the length of the game.	Indoors = 0, None = 1, 2 = Calm, 3 = Varies, 4 = Out to CF, 5 = RF to LF, 6 = Out to RF, 7 = Out to LF, 9 = In from LF, 10 = In from LF, 11 = In from RF	0-11
wind_speed	The speed of the wind, measured at the start of the game, in miles per hour.	integer, in mph	0-28
Car_bat_H	The number of times the batter has hit the ball into fair territory in his career.	integer	0-5050
Car_bat_G	The number of games the batter has played in his career.	integer	1-1506
Car_bat_PA	The number of plate appearances in the batter's career.	integer	1-6626
Car_bat_HBP	The number of times the batter has been hit by a pitch in his career.	integer	0-115
Car_bat_BB	The number of times the batter has been walked in his career.	integer	0-958
Car_bat_K	The number of times the batter has struck out in his career.	integer	0-1660

Car_bat_Swings	The number of times the batter has swung at a pitch in his career.	integer	1-12817
Car_bat_Whiffs	The number of times the batter has swung and missed a pitch in his career.	integer	0-341
Car_pit_Whiffs	The number of times the pitcher has thrown a pitch that the batter swung and missed at in his career.	integer	0-403
Car_pit_Swings	The number of times the pitcher has thrown a pitch that the batter swung at in his career.	integer	2-16912
Car_pit_IP	The number of innings pitched in the pitcher's career.	integer	0.6666667 through 2859
Car_pit_K	The number of strikeouts the pitcher has thrown in his career.	integer	0-2174
Car_pit_BB	The number of walks the pitcher has allowed in his career.	integer	0-702
Car_pit_HR	The number of homeruns the pitcher has allowed in his career.	integer	0-287
Car_pit_G	The number of games a pitcher has played in his career.	integer	1-692
Car_pit_H	The number of hits a pitcher has allowed in his career.	integer	2-5822

Car_pit_BF	The number of batters faced by a pitched in his career.	integer	2-8577
Biseason_bat_H	The number of times the batter has hit the ball into fair territory in the past two years.	integer	0-948
Biseason_bat_G	The number of games the batter has played in the last two years.	integer	1-270
Biseason_bat_PA	The number of plate appearances the batter has had in the past two years.	integer	1-1268
Biseason_bat_HBP	The number of times the batter has been hit by a pitch in the past two years.	integer	0-39
Biseason_bat_BB	The number of walks the batter has had in the last two years.	integer	0-186
Biseason_bat_K	The number of times the batter has struck out in the last two years.	integer	0-358
Biseason_bat_Swings	The number of times the batter has swung at a pitch in the last two years.	integer	1-2376
Biseason_bat_Whiffs	The number of times the batter has swung and missed a pitch to strike out in the past two years.	integer	0-57
Biseason_pit_Whiffs	The number of times the pitcher has thrown a pitch that the batter	integer	0-133

	swung at and missed in the past two years.		
Biseason_pit_Swings	The number of pitched the pitcher has thrown that the batter swung at in the past two years.	integer	0-3931
Biseason_pit_IP	The number of innings pitched in the pitcher's past two years.	continuous	0.6666667 through 662.3333
Biseason_pit_K	The number of strikeouts the pitcher has thrown in the past two years.	integer	0-741
Biseason_pit_BB	The number of walks the pitcher has allowed in the past two years.	integer	0-158
Biseason_pit_HR	The number of home run the pitcher has allowed in the past two years.	integer	0-64
Biseason_pit_G	The number of games the pitcher has played in the past two years.	integer	1-128
Biseason_pit_H	The number of contacts into fair territory the pitcher has allowed in the past two years.	integer	0-1166
Biseason_pit_BF	The number of batters faced by the pitcher in the past two years.	integer	2-1987
car_bat_h_per_game	The average number of contacts into fair territory per game the batter has had in his career.	continuous	0-9

car_bat_hits	The number of contacts into fair territory the batter has had in his career.	integer	0-5050
car_bat_avg_pa	The average number of plate appearances the batter has had per game in his career.	continuous	1-11.5
car_bat_ba	The batter's career batting average for contacts into fair territory.	continuous	0-1
car_bat_pitch_prop_zone(1-30)	The proportion of pitches the batter has been thrown per zone out of all of the pitches the batter has received in his career.	continuous	N/A
car_bat_swing_prop_zone(1-30)	The proportion of swings the batter has taken per zone out of all of the pitches he has received in his career.	continuous	N/A
car_bat_whiff_prop_zone(1-30)	The proportion of swing and misses the batter has had per zone out of all of the pitches he has received in his career.	continuous	N/A
car_bat_k_bb_prop_zone(1-30)	The proportion of times the batter has either walked or struck out per zone out of all of the pitches he has received in his career.	continuous	N/A

car_bat_hit_prop_pitch(0-16)	The proportion of times the batter has hit the ball into fair territory per pitch type out of all of the pitches he has received in his career.	continuous	N/A
car_bat_nohit_prop_pitch(0-16)	The number of times the batter has failed to hit the ball into fair territory per pitch type out of all of the pitches he has received in his career.	continuous	N/A
car_bat_ba_prop_zone(1-30)	The batter's hitting average (total hits divided by total plate appearances) per zone, out of all of the pitches he has received in his career.	continuous	N/A
biseason_bat_h_per_avg_pa	The number of times the batter has hit the ball into fair territory per his average number of plate appearances per game in the last two years.	continuous	0 through 216.4265
biseason_bat_hits	The number of times the batter has hit the ball into fair territory in the last two years.	integer	0-948
biseason_bat_avg_pa	The average number of plate appearances the batter has had per game in the last two years.	continuous	1 through 11.5
biseason_bat_ba	The batter's hitting average (total hits into fair territory divided by	continuous	0 through 1

	total plate appearances) in the last two years.		
bi_bat_pitch_prop_zone(1-30)	The proportion of pitches the batter has received per zone out of all the pitches he has received in the last two years.	continuous	N/A
bi_bat_swing_prop_zone(1-30)	The proportion of swings the batter has taken per zone out of all the pitches he has received in the last two years.	continuous	N/A
bi_bat_whiff_prop_zone(1-30)	The proportion of swings and misses the batter has had per zone out of all the pitches he has received in the last two years.	continuous	N/A
bi_bat_k_bb_prop_zone(1-30)	The proportion of strikeouts or walks for the batter per zone out of all the pitches he has received in the last two years.	continuous	N/A
bi_bat_hit_prop_pitch(0-16)	The proportion of hits into fair territory the batter has had per pitch type out of all the pitches he has received in the last two years.	continuous	N/A
bi_bat_nohit_prop_pitch(0-16)	The proportion of times the batter has failed to hit the ball per pitch type out of all the	continuous	N/A

	pitches he has received in the last two years.		
bi_bat_ba_prop_zone(1-30)	The batter's hitting average (total hits into fair territory divided by total plate appearances) per zone out of all the pitches he has received in the last two years.	continuous	N/A
car_pit_FIP	The pitcher's Fielding Independent Pitching for his career.	continuous	0.7098182 through 27.10982
car_pit_h9	The total number of hits allowed by the pitcher per total innings pitched divided by 9 (i.e. per every 9 innings pitched) in his career.	continuous	9.346154 through 27
car_pit_noh9	The total number of times the pitcher has prevented the batter from hitting the ball per total innings pitched divided by 9 (i.e. per every 9 innings pitched) in his career.	continuous	0 through 17.65385
car_pit_k_bb	The pitcher's overall strikeout to walk ratio for his career.	continuous	0 through Infinity
car_pit_hits	The total number of hits allowed by the pitcher in his career.	integer	2-5822
car_pit_avg_ip	The pitcher's average number of innings pitched per game in his career.	continuous	0.6666667 through 22.66667

car_pit_ba	The batting average against the pitcher over his entire career.	continuous	0.3461538 through 1
car_pit_ba_prop_zone(1-30)	The hitting average against the pitcher per zone out of all of the pitches the pitcher has thrown over his entire career.	continuous	N/A
car_pit_pitches_prop_zone(1-30)	The proportion of pitches thrown per zone out of all of the pitches thrown by the pitcher over his entire career.	continuous	N/A
car_pit_swings_prop_zone(1-30)	The proportion of swings against the pitcher per zone out of all the pitches thrown by the pitcher over his career.	continuous	N/A
car_pit_whiffs_prop_zone(1-30)	The proportion of swing and misses against the pitcher per zone out of all the pitches thrown by the pitcher over his career.	continuous	N/A
car_pit_hits_prop_zone(1-30)	The proportion of hits into fair territory allowed by the pitcher per zone out of all his pitches thrown in his career.	continuous	N/A
car_pit_nohits_prop_zone(1-30)	The proportion of hits into fair territory prevented by the pitcher per zone out of all of his	continuous	N/A

	pitches thrown in his career.		
biseason_pit_FIP	The pitcher's Fielding Independent Pitching over the last two years.	continuous	0.7595 through 63.1595
biseason_pit_h9	The number of hits allowed by the pitcher per total innings pitched divided by 9 (i.e. per every 9 innings pitched) in the last two years)	continuous	0-27
biseason_pit_noh9	The number of hits prevented by the pitcher per total innings pitched divided by 9 (i.e. per every 9 innings pitched) in the past two years.	continuous	0-27
biseason_pit_k_bb	The pitcher's strikeout to walk ratio over the last two years.	continuous	0 through Infinity
biseason_pit_hits	The number of hits allowed by the pitcher over the last two years.	integer	0-1166
biseason_pit_avg_ip	The average number of innings pitched per game for the pitcher over the last two years.	continuous	0.375 through 25
biseason_pit_ba	The batting average against the pitcher over the last two years.	continuous	0 through 1
bi_pit_ba_prop_zone(1-30)	The hitting average against the pitcher per zone out of his pitches thrown over the last two years.	continuous	N/A

bi_pit_pitches_prop_zone(1-30)	The proportion of pitches thrown by the pitcher per zone out of all his pitches thrown over the last two years.	continuous	N/A
bi_pit_swings_prop_zone(1-30)	The proportion of swings against the pitcher per zone out of all his pitches thrown over the last two years.	continuous	N/A
bi_pit_whiffs_prop_zone(1-30)	THe proportion of swing and misses against the pitcher per zone out of all his pitches thrown over the last two years.	continuous	N/A
bi_pit_hits_prop_zone(1-30)	The proportion of hits allowed by the pitcher per zone out of all his pitches thrown over the last two years.	continuous	N/A
bi_pit_nohits_prop_zone(1-30)	The proportion of hits prevented by the pitcher per zone out of all his pitches thrown over the last two years.	continuous	N/A
mat_car_pitch_prop_zone(1-30)	The proportion of pitches the batter received per zone out of all the pitches the batter received in his career times the proportion of pitches the pitcher has thrown per zone out of all the pitches he has thrown in his career.	continuous	N/A

mat_car_swing_prop_zone(1-30)	The proportion of swings the batter has taken per zone out of all the pitches he has received in his career times the proportion of swings against the pitcher per zone out of all of the pitches he has thrown in his career.	continuous	N/A
mat_car_whiff_prop_zone(1-30)	The proportion of swing and misses the batter has had per zone out of all of the pitches he has received in his career times the proportion of swing and misses against the pitcher per zone out of all the pitches he has thrown in his career.	continuous	N/A
mat_car_k_bb_prop_zone(1-30)	The proportion of strikeouts and walks for the batter per zone out of all the pitches he has received in his career time the proportion of strikeouts and walks by the pitcher per zone out of all of the pitches he has thrown in his career.	continuous	N/A
mat_car_ba_prop_zone(1-30)	The batter's hitting average per zone out of all the pitches he has received in his career times the hitting average against the pitcher out	continuous	N/A

	of all the pitches he has thrown in his career.		
mat_bi_pitch_prop_zone(1-30)	The proportion of pitches the batter received per zone out of all the pitches the batter received in the past two years times the proportion of pitches the pitcher has thrown per zone out of all the pitches he has thrown in the past two years.	continuous	N/A
mat_bi_swing_prop_zone(1-30)	The proportion of swings the batter has taken per zone out of all the pitches he has received in the past two years times the proportion of swings against the pitcher per zone out of all of the pitches he has thrown in the past two years.	continuous	N/A
mat_bi_whiff_prop_zone(1-30)	The proportion of swing and misses the batter has had per zone out of all of the pitches he has received in the past two years times the proportion of swing and misses against the pitcher per zone out of all the pitches he has thrown in the past two years.	continuous	N/A

mat_bi_k_bb_prop_zone(1-30)	The proportion of strikeouts and walks for the batter per zone out of all the pitches he has received in the past two years time the proportion of strikeouts and walks by the pitcher per zone out of all of the pitches he has thrown in the past two years.	continuous	N/A
mat_bi_ba_prop_zone(1-30)	The batter's hitting average per zone out of all the pitches he has received in the past two years times the hitting average against the pitcher out of all the pitches he has thrown in the past two years.	continuous	N/A