In [32]:

```
import pybaseball as pyb
from pybaseball import statcast, pitching_stats, playerid_lookup, stat
import numpy as np
import math
import pandas as pd
import glob
import os
import re
import unicodedata
from datetime import datetime
from itertools import groupby
from operator import itemgetter
from sklearn.preprocessing import OneHotEncoder
```

First, load in DF from data_cleaning_notebook_1 'cleaning_filtered_df.csv'

Drop all info prior to 2008.

Then rename 'Year' to 'season'.

In [23]:
```
cleaning_filtered_df = pd.read_csv('~/Documents/Flatiron/Project_5_/da
cleaning_filtered_df
```

Out[23]:

	Name	Age	Year	Throws	IP	G	GS	CG	SHO	sDR	Career Start	Career End	Inact Ye
0	ed acosta	28	1972	1	89.0	46	2	0	0	1	1972	1972	
1	doyle alexander	21	1972	1	106.1	35	9	2	2	3	1972	1989	
2	lloyd allen	22	1972	1	85.1	42	6	0	0	5	1972	1975	
3	steve arlin	26	1972	1	250.0	38	37	12	3	22	1972	1974	
4	stan bahnsen	27	1972	1	252.1	43	41	5	1	32	1972	1981	[19 19
...	...	...	...	...	...	...	...	...	...	...	...	...	
16566	brandon woodruff	30	2023	1	67.0	11	11	1	1	0	2017	2023	
16567	kyle wright	27	2023	1	31.0	9	7	0	0	0	2019	2023	
16568	ryan yarbrough	31	2023	0	89.2	25	9	0	0	0	2018	2023	
16571	rob zastryzny	31	2023	0	20.2	21	1	0	0	0	2016	2023	[20 20 20 20 20 20
16572	angel zerpa	23	2023	0	42.2	15	3	0	0	1	2021	2023	

15060 rows × 13 columns

In [24]:
```
1 cleaning_filtered_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 15060 entries, 0 to 16572
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Name            15060 non-null  object
 1   Age             15060 non-null  int64
 2   Year            15060 non-null  int64
 3   Throws          15060 non-null  int64
 4   IP              15060 non-null  float64
 5   G               15060 non-null  int64
 6   GS              15060 non-null  int64
 7   CG              15060 non-null  int64
 8   SHO             15060 non-null  int64
 9   sDR             15060 non-null  int64
 10  Career Start    15060 non-null  int64
 11  Career End      15060 non-null  int64
 12  Inactive Years  15060 non-null  object
dtypes: float64(1), int64(10), object(2)
memory usage: 1.6+ MB
```

In [25]:
```
1 cleaning_filtered_df = cleaning_filtered_df[cleaning_filtered_df['Year
```

In [26]:
```
1 cleaning_filtered_df
```

Out[26]:

	Name	Age	Year	Throws	IP	G	GS	CG	SHO	sDR	Career Start	Career End	Inactiv Yea
10745	alfredo aceves	25	2008	1	30.0	6	4	0	0	0	2008	2013	[201 201
10746	nick adenhart	21	2008	1	12.0	3	3	0	0	0	2008	2009	
10747	matt albers	25	2008	1	49.0	28	3	0	0	2	2006	2016	[200 201 201 201 201 201 201
10748	alberto arias	24	2008	1	21.2	15	2	0	0	0	2008	2008	
10749	alberto arias	24	2008	1	8.0	3	2	0	0	0	2008	2008	
...	...	...	...	...	...	...	...	...	...	...	...	...	
16566	brandon woodruff	30	2023	1	67.0	11	11	1	1	0	2017	2023	
16567	kyle wright	27	2023	1	31.0	9	7	0	0	0	2019	2023	
16568	ryan yarbrough	31	2023	0	89.2	25	9	0	0	0	2018	2023	
16571	rob zastryzny	31	2023	0	20.2	21	1	0	0	0	2016	2023	[201 201 201 202 202 202
16572	angel zerpa	23	2023	0	42.2	15	3	0	0	1	2021	2023	

5305 rows × 13 columns

In [27]:
```
1 cleaning_filtered_df = cleaning_filtered_df.rename(columns={'Year': 's
```

Will merge cleaning_filtered_df with yearly DF later.

Now, load in data by year.

Start with 2008.

In [9]:
```
1 all_2008_stats_df = pd.read_csv('~/Documents/Flatiron/Project_5_/data/
```

In [10]:

```
1 all_2008_stats_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 459185 entries, 0 to 459184
Data columns (total 92 columns):
 #   Column                          Non-Null Count   Dtype  
---  ------                          --------------   -----  
 0   pitch_type                      437522 non-null  object 
 1   game_date                       459185 non-null  object 
 2   release_speed                   437503 non-null  float64
 3   release_pos_x                   437492 non-null  float64
 4   release_pos_z                   437492 non-null  float64
 5   player_name                     459185 non-null  object 
 6   batter                          459185 non-null  int64  
 7   pitcher                         459185 non-null  int64  
 8   events                          122796 non-null  object 
 9   description                     459185 non-null  object 
 10  spin_dir                        0 non-null       float64
 11  spin_rate_deprecated            0 non-null       float64
 12  break_angle_deprecated          0 non-null       float64
 13  break_length_deprecated         0 non-null       float64
 14  zone                            437503 non-null  float64
 15  des                             459184 non-null  object 
 16  game_type                       459185 non-null  object 
 17  stand                           459185 non-null  object 
 18  p_throws                        459185 non-null  object 
 19  home_team                       459185 non-null  object 
 20  away_team                       459185 non-null  object 
 21  type                            459185 non-null  object 
 22  hit_location                    108130 non-null  float64
 23  bb_type                         90752 non-null   object 
 24  balls                           459185 non-null  int64  
 25  strikes                         459185 non-null  int64  
 26  game_year                       459185 non-null  int64  
 27  pfx_x                           437492 non-null  float64
 28  pfx_z                           437492 non-null  float64
 29  plate_x                         437503 non-null  float64
 30  plate_z                         437503 non-null  float64
 31  on_3b                           43982 non-null   float64
 32  on_2b                           87192 non-null   float64
 33  on_1b                           138014 non-null  float64
 34  outs_when_up                    459185 non-null  int64  
 35  inning                          459185 non-null  int64  
 36  inning_topbot                   459185 non-null  object 
 37  hc_x                            82377 non-null   float64
 38  hc_y                            82377 non-null   float64
 39  tfs_deprecated                  0 non-null       float64
 40  tfs_zulu_deprecated             0 non-null       float64
 41  fielder_2                       459185 non-null  int64  
 42  umpire                          0 non-null       float64
 43  sv_id                           437521 non-null  object 
 44  vx0                             437492 non-null  float64
 45  vy0                             437492 non-null  float64
 46  vz0                             437492 non-null  float64
 47  ax                              437503 non-null  float64
 48  ay                              437503 non-null  float64
 49  az                              437503 non-null  float64
 50  sz_top                          437503 non-null  float64
 51  sz_bot                          437503 non-null  float64
```

```
 52  hit_distance_sc                  0 non-null       float64
 53  launch_speed                     0 non-null       float64
 54  launch_angle                     0 non-null       float64
 55  effective_speed                  0 non-null       float64
 56  release_spin_rate                0 non-null       float64
 57  release_extension                0 non-null       float64
 58  game_pk                          459185 non-null  int64
 59  pitcher.1                        459185 non-null  int64
 60  fielder_2.1                      459185 non-null  int64
 61  fielder_3                        459185 non-null  int64
 62  fielder_4                        459185 non-null  int64
 63  fielder_5                        459185 non-null  int64
 64  fielder_6                        459185 non-null  int64
 65  fielder_7                        459185 non-null  int64
 66  fielder_8                        459185 non-null  int64
 67  fielder_9                        459185 non-null  int64
 68  release_pos_y                    437492 non-null  float64
 69  estimated_ba_using_speedangle    0 non-null       float64
 70  estimated_woba_using_speedangle  0 non-null       float64
 71  woba_value                       122797 non-null  float64
 72  woba_denom                       0 non-null       float64
 73  babip_value                      122797 non-null  float64
 74  iso_value                        122797 non-null  float64
 75  launch_speed_angle               0 non-null       float64
 76  at_bat_number                    459185 non-null  int64
 77  pitch_number                     459185 non-null  int64
 78  pitch_name                       437522 non-null  object
 79  home_score                       459185 non-null  int64
 80  away_score                       459185 non-null  int64
 81  bat_score                        459185 non-null  int64
 82  fld_score                        459185 non-null  int64
 83  post_away_score                  459185 non-null  int64
 84  post_home_score                  459185 non-null  int64
 85  post_bat_score                   459185 non-null  int64
 86  post_fld_score                   459185 non-null  int64
 87  if_fielding_alignment            0 non-null       float64
 88  of_fielding_alignment            0 non-null       float64
 89  spin_axis                        0 non-null       float64
 90  delta_home_win_exp               459185 non-null  float64
 91  delta_run_exp                    449456 non-null  float64
dtypes: float64(48), int64(28), object(16)
memory usage: 325.8+ MB
```

In [11]:

```
all_2008_stats_df.drop(columns=['batter', 'events', 'description', 'zo
                                'des', 'game_type', 'stand', 'home_tea
                                'away_team', 'type', 'hit_location', '
                                'balls', 'strikes', 'pfx_x', 'spin_dir
                                'pfx_z', 'plate_x', 'plate_z', 'on_3b'
                                'on_2b', 'on_1b', 'outs_when_up', 'inn
                                'inning_topbot', 'hc_x', 'hc_y', 'fiel
                                'umpire', 'sv_id', 'hit_distance_sc',
                                 'sz_bot', 'launch_speed', 'launch_ang
                                'pitcher.1', 'fielder_2.1', 'fielder_3
                                'fielder_5', 'fielder_6', 'fielder_7',
                                'fielder_9', 'estimated_ba_using_speed
                                'estimated_woba_using_speedangle', 'ba
                                'launch_speed_angle', 'woba_value', 'w
                                'at_bat_number', 'pitch_number', 'home
                                'bat_score', 'fld_score', 'post_home_s
                                'post_fld_score', 'post_away_score', '
                                'of_fielding_alignment', 'delta_home_w
                                'delta_run_exp', 'spin_rate_deprecated
                               'break_length_deprecated', 'tfs_deprecat
                               'spin_axis', 'effective_speed', 'releas
all_2008_stats_df.head()
```

Out[11]:

	pitch_type	game_date	release_speed	release_pos_x	release_pos_z	player_name	pitche
0	FF	2008-09-28	94.4	2.17	6.08	Rhodes, Arthur	12112
1	FF	2008-09-28	93.0	2.21	6.03	Rhodes, Arthur	12112
2	NaN	2008-09-26	NaN	NaN	NaN	Rhodes, Arthur	12112
3	NaN	2008-09-26	NaN	NaN	NaN	Rhodes, Arthur	12112
4	FF	2008-09-22	92.2	1.87	6.72	Rhodes, Arthur	12112

In [12]:
```
1 all_2008_stats_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 459185 entries, 0 to 459184
Data columns (total 17 columns):
 #   Column         Non-Null Count   Dtype  
---  ------         --------------   -----  
 0   pitch_type     437522 non-null  object 
 1   game_date      459185 non-null  object 
 2   release_speed  437503 non-null  float64
 3   release_pos_x  437492 non-null  float64
 4   release_pos_z  437492 non-null  float64
 5   player_name    459185 non-null  object 
 6   pitcher        459185 non-null  int64  
 7   p_throws       459185 non-null  object 
 8   game_year      459185 non-null  int64  
 9   vx0            437492 non-null  float64
 10  vy0            437492 non-null  float64
 11  vz0            437492 non-null  float64
 12  ax             437503 non-null  float64
 13  ay             437503 non-null  float64
 14  az             437503 non-null  float64
 15  release_pos_y  437492 non-null  float64
 16  pitch_name     437522 non-null  object 
dtypes: float64(10), int64(2), object(5)
memory usage: 63.1+ MB
```

In [13]:
```
1 all_2008_stats_df = all_2008_stats_df.dropna(axis=0)
```

In [14]:
```
1 all_2008_stats_df.reset_index(inplace=True)
```

In [15]:
```
1 all_2008_stats_df.drop('index', axis=1)
```

Out[15]:

	pitch_type	game_date	release_speed	release_pos_x	release_pos_z	player_name
0	FF	2008-09-28	94.4	2.17	6.08	Rhodes, Arthur
1	FF	2008-09-28	93.0	2.21	6.03	Rhodes, Arthur
2	FF	2008-09-22	92.2	1.87	6.72	Rhodes, Arthur
3	SL	2008-09-22	83.5	1.67	6.77	Rhodes, Arthur
4	FF	2008-09-22	92.7	1.94	6.68	Rhodes, Arthur
...	...	...	...	...	...	...
437487	SI	2008-04-05	91.7	-1.14	6.77	Wainwright, Adam
437488	SI	2008-04-05	91.2	-1.23	6.71	Wainwright, Adam
437489	SI	2008-04-05	91.5	-0.97	6.65	Wainwright, Adam
437490	CU	2008-04-05	72.3	-1.15	6.93	Wainwright, Adam
437491	SI	2008-04-05	90.0	-1.01	6.71	Wainwright, Adam

437492 rows × 17 columns

In [16]:

```
# Group by 'game_date' and 'pitcher' to calculate the total pitches
total_pitches = all_2008_stats_df.groupby(['game_date', 'pitcher', 'pl

# Group by 'game_date', 'pitcher', and 'pitch_type' to calculate the s
total_pitches_by_type = all_2008_stats_df.groupby(['game_date', 'pitch

# Calculate averages of the specified metrics for each pitch type, gro
avg_metrics = all_2008_stats_df.groupby(['game_date', 'pitcher', 'play
    'release_speed': 'mean',
    'release_pos_x': 'mean',
    'release_pos_z': 'mean',
    'vx0': 'mean',
    'vy0': 'mean',
    'vz0': 'mean',
    'ax': 'mean',
    'ay': 'mean',
    'az': 'mean',
    'release_pos_y': 'mean',
}).reset_index()

grouped_2008_df = total_pitches.merge(total_pitches_by_type, on=['game
grouped_2008_df = grouped_2008_df.merge(avg_metrics, on=['game_date',

grouped_2008_df
```

Out[16]:

	game_date	pitcher	player_name	total_pitches	pitch_type	count_by_pitch_type	rele
0	2008-03-28	112526	Colon, Bartolo	39	FF	10	
1	2008-03-28	112526	Colon, Bartolo	39	SI	17	
2	2008-03-28	112526	Colon, Bartolo	39	SL	12	
3	2008-03-28	118120	Maddux, Greg	1	SI	1	
4	2008-03-28	121556	Rusch, Glendon	8	FF	4	
...	...	...	...	...	...	...	
28419	2008-09-30	448147	Blackburn, Nick	89	CH	3	
28420	2008-09-30	448147	Blackburn, Nick	89	CU	4	
28421	2008-09-30	448147	Blackburn, Nick	89	FC	21	
28422	2008-09-30	448147	Blackburn, Nick	89	IN	4	
28423	2008-09-30	448147	Blackburn, Nick	89	SI	57	

28424 rows × 16 columns

In [17]:

```
grouped_2008_df['game_date'] = pd.to_datetime(grouped_2008_df['game_da
grouped_2008_df['season'] = grouped_2008_df['game_date'].dt.year

# Step 1: Season Total Pitches
season_total_pitches = grouped_2008_df.groupby(['pitcher', 'player_nam

# Step 2: Season Total by Pitch Type
season_total_by_pitch_type = grouped_2008_df.groupby(['pitcher', 'play

# Weighted Averages Calculation Setup
weighted_avg_columns = ['release_speed', 'release_pos_x', 'release_pos
for col in weighted_avg_columns:
    grouped_2008_df[f'{col}_product'] = grouped_2008_df[col] * grouped

weighted_avg_aggregations = {f'{col}_product': 'sum' for col in weight
weighted_avg_aggregations['count_by_pitch_type'] = 'sum'

# Aggregate for weighted averages
weighted_avg_df = grouped_2008_df.groupby(['pitcher', 'player_name', '

# Calculate weighted averages
for col in weighted_avg_columns:
    weighted_avg_df[f'{col}_weighted_avg'] = weighted_avg_df[f'{col}_p

# Cleanup
weighted_avg_df.drop(columns=[f'{col}_product' for col in weighted_avg

# Merge season totals and weighted averages
final_2008_df = pd.merge(season_total_pitches, season_total_by_pitch_t
final_2008_df = pd.merge(final_2008_df, weighted_avg_df, on=['pitcher'

final_2008_df.head()
```

Out[17]:

	pitcher	player_name	season	season_total_pitches	pitch_type	season_total_count_by_pit
0	110683	Batista, Miguel	2008	10287	CH	
1	110683	Batista, Miguel	2008	10287	CU	
2	110683	Batista, Miguel	2008	10287	FC	
3	110683	Batista, Miguel	2008	10287	FF	
4	110683	Batista, Miguel	2008	10287	IN	

In [18]:
```
final_2008_df['player_name'] = final_2008_df['player_name'].str.lower(
final_2008_df
```

Out[18]:

	pitcher	player_name	season	season_total_pitches	pitch_type	season_total_count_by
0	110683	batista, miguel	2008	10287	CH	
1	110683	batista, miguel	2008	10287	CU	
2	110683	batista, miguel	2008	10287	FC	
3	110683	batista, miguel	2008	10287	FF	
4	110683	batista, miguel	2008	10287	IN	
...	...	...	...	...	...	
1574	502188	samardzija, jeff	2008	1820	FS	
1575	502188	samardzija, jeff	2008	1820	IN	
1576	502188	samardzija, jeff	2008	1820	PO	
1577	502188	samardzija, jeff	2008	1820	SI	
1578	502188	samardzija, jeff	2008	1820	SL	

1579 rows × 17 columns

In [19]:

```
1 print(final_2008_df['player_name'].unique())
```

```
['batista, miguel' 'brocail, doug' 'byrd, paul' 'carpenter, chris'
 'colon, bartolo' 'dessens, elmer' 'estes, shawn' 'eyre, scott'
 'glavine, tom' 'gordon, tom' 'hampton, mike' 'hawkins, latroy'
 'hernandez, livan' 'isringhausen, jason' 'lieber, jon' 'loaiza, esteban'
 'lowe, derek' 'maddux, greg' 'martinez, pedro' 'mercker, kent'
 'millwood, kevin' 'moehler, brian' 'morris, matt' 'moyer, jamie'
 'mussina, mike' 'nomo, hideo' 'oliver, darren' 'park, chan ho'
 'pettitte, andy' 'reyes, dennys' 'rhodes, arthur' 'rogers, kenny'
 'rusch, glendon' 'smoltz, john' 'springer, russ' 'sturtze, tanyon'
 'suppan, jeff' 'tavarez, julian' 'tomko, brett' 'torres, salomon'
 'trachsel, steve' 'villone, ron' 'wakefield, tim' 'weathers, david'
 'williams, woody' 'wright, jamey' 'looper, braden' 'washburn, jarrod'
 'ponson, sidney' 'dempster, ryan' 'elarton, scott' 'wood, kerry'
 'vazquez, javier' 'pavano, carl' 'perez, odalis' 'nunez, vladimir'
 'halladay, roy' 'schoeneweis, scott' 'wolf, randy' 'figueroa, nelson'
 'redman, mark' 'nathan, joe' 'davis, doug' 'marquis, jason'
 'moss, damian' 'fogg, josh' 'wells, kip' 'burnett, a.j.' 'armas, tony'
 'lilly, ted' 'westbrook, jake' 'mulder, mark' 'glover, gary'
 'penny, brad' 'franklin, ryan' 'zito, barry' 'hudson, tim'
 'padilla, vicente' 'meche, gil' 'durbin, chad' 'downs, scott'
 'chacon, shawn' 'santana, johan' 'arroyo, bronson' 'benoit, joaquín'
 'beckett, josh' 'belisle, matt' 'garland, jon' 'buehrle, mark'
 'sabathia, cc' 'vargas, claudio' 'sheets, ben' 'eaton, adam'
 'dickey, r.a.' 'pineiro, joel' 'affeldt, jeremy' 'lohse, kyle'
 'duckworth, brandon' 'cook, aaron' 'oswalt, roy' 'redding, tim'
 'silva, carlos' 'ramirez, horacio' 'moseley, dustin' 'fossum, casey'
 'zambrano, carlos' 'jennings, jason' 'lackey, john' 'de la rosa, jorge'
 'backe, brandon' 'bedard, erik' 'sosa, jorge' 'myers, brett'
 'peavy, jake' 'harang, aaron' 'pérez, oliver' 'saarloos, kirk'
 'hernandez, runelvys' 'hendrickson, mark' 'robertson, nate'
 'davis, jason' 'tallet, brian' 'guthrie, jeremy' 'wang, chien-ming'
 'gobble, jimmy' 'wellemeyer, todd' 'mcclung, seth' 'kuo, hung-chih'
 'cabrera, daniel' 'webb, brandon' 'carrasco, d.j.' 'ledezma, wilfredo'
 'contreras, jose' 'wainwright, adam' 'bonser, boof' 'bonderman, jeremy'
 'greinke, zack' 'harden, rich' 'floyd, gavin' 'willis, dontrelle'
 'haren, dan' 'hill, shawn' 'jackson, edwin' 'maine, john'
 'santana, ervin' 'correia, kevin' 'gaudin, chad' 'simon, alfredo'
 'blanton, joe' 'snell, ian' 'mcgowan, dustin' 'germano, justin'
 'maholm, paul' 'cain, matt' 'hamels, cole' 'kazmir, scott' 'young, chri
s'
 'danks, john' 'thompson, brad' 'hernandez, roberto' 'francis, jeff'
 'hernández, félix' 'petit, yusmeiro' 'bush, dave' 'loe, kameron'
 'verlander, justin' 'liriano, francisco' 'tejeda, robinson'
 'saunders, joe' 'jiménez, ubaldo' 'hammel, jason' 'burres, brian'
 'rodriguez, wandy' 'mendoza, luis' 'sánchez, aníbal' 'davies, kyle'
 'rowland-smith, ryan' 'reyes, anthony' 'duke, zach' 'olsen, scott'
 'mccarthy, brandon' 'miner, zach' 'niemann, jeff' 'karstens, jeff'
 'laffey, aaron' 'feldman, scott' 'nolasco, ricky' 'marshall, sean'
 'stults, eric' 'chavez, jesse' 'eveland, dana' 'litsch, jesse'
 'bannister, brian' 'blackburn, nick' 'parra, manny' 'hill, rich'
 'shields, james' 'wells, randy' 'garcía, jaime' 'volquez, edinson'
 'weaver, jered' 'wilson, c.j.' 'galarraga, armando' 'billingsley, chad'
 'gallardo, yovani' 'marcum, shaun' 'owings, micah' 'lester, jon'
 'kendrick, kyle' 'gorzelanny, tom' 'kennedy, ian' 'miller, andrew'
 'leblanc, wade' 'scherzer, max' 'lincecum, tim' 'buchholz, clay'
 'morrow, brandon' 'richard, clayton' 'villanueva, carlos' 'hensley, cla
y'
```

```
'ohlendorf, ross' 'price, david' 'sánchez, jonathan' 'cueto, johnny'
'bergmann, jason' 'bailey, homer' 'harrison, matt' 'jurrjens, jair'
'reyes, jo-jo' 'happ, j.a.' 'sonnanstine, andy' 'volstad, chris'
'lannan, john' 'slowey, kevin' 'humber, philip' 'hochevar, luke'
'pelfrey, mike' 'sowers, jeremy' 'braden, dallas' 'gonzález, gio'
'hughes, phil' 'estrada, marco' 'morales, franklin' 'masterson, justin'
'niese, jonathon' 'kershaw, clayton' 'hunter, tommy' 'garza, matt'
'kuroda, hiroki' 'matsuzaka, daisuke' 'samardzija, jeff']
```

In [20]:

```
def remove_accents(input_str):
    nfkd_form = unicodedata.normalize('NFKD', input_str)
    return "".join([c for c in nfkd_form if not unicodedata.combining(

def clean_name(name):
    name = name.lower()
    name = remove_accents(name)
    name = re.sub(r'[-.]', '', name)
    name = re.sub(r'\s+', ' ', name).strip()
    return name

final_2008_df['player_name'] = final_2008_df['player_name'].apply(clea
```

In [21]:

```
# Convert 'player_name' from "last name, first name" to "first name la
final_2008_df['Name'] = final_2008_df['player_name'].apply(lambda x: '
```

Finally!

Now merge the two DF on 'Name' and 'season'.

In [28]:

```
# Now, you can perform the merge using 'Name' and 'season' as the keys
better_2008_df = pd.merge(final_2008_df,
                          cleaning_filtered_df[['Name', 'season', 'Age
                          on=['Name', 'season'],
                          how='left')
```

In [29]:
```
1 better_2008_df
```

Out[29]:

	pitcher	player_name	season	season_total_pitches	pitch_type	season_total_count_by
0	110683	batista, miguel	2008	10287	CH	
1	110683	batista, miguel	2008	10287	CU	
2	110683	batista, miguel	2008	10287	FC	
3	110683	batista, miguel	2008	10287	FF	
4	110683	batista, miguel	2008	10287	IN	
...	...	...	...	...	...	
1621	502188	samardzija, jeff	2008	1820	FS	
1622	502188	samardzija, jeff	2008	1820	IN	
1623	502188	samardzija, jeff	2008	1820	PO	
1624	502188	samardzija, jeff	2008	1820	SI	
1625	502188	samardzija, jeff	2008	1820	SL	

1626 rows × 19 columns

In [30]:
```
1 better_2008_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1626 entries, 0 to 1625
Data columns (total 19 columns):
 #   Column                            Non-Null Count  Dtype
---  ------                            --------------  -----
 0   pitcher                           1626 non-null   int64
 1   player_name                       1626 non-null   object
 2   season                            1626 non-null   int32
 3   season_total_pitches              1626 non-null   int64
 4   pitch_type                        1626 non-null   object
 5   season_total_count_by_pitch_type  1626 non-null   int64
 6   count_by_pitch_type               1626 non-null   int64
 7   release_speed_weighted_avg        1626 non-null   float64
 8   release_pos_x_weighted_avg        1626 non-null   float64
 9   release_pos_z_weighted_avg        1626 non-null   float64
 10  vx0_weighted_avg                  1626 non-null   float64
 11  vy0_weighted_avg                  1626 non-null   float64
 12  vz0_weighted_avg                  1626 non-null   float64
 13  ax_weighted_avg                   1626 non-null   float64
 14  ay_weighted_avg                   1626 non-null   float64
 15  az_weighted_avg                   1626 non-null   float64
 16  release_pos_y_weighted_avg        1626 non-null   float64
 17  Name                              1626 non-null   object
 18  Age                               1358 non-null   float64
dtypes: float64(11), int32(1), int64(4), object(3)
memory usage: 235.1+ KB
```

In [31]:
```
1 """
2 better_2008_df.to_csv('data/better_2008_df.csv')
3 """
```

Out[31]:
```
"\nbetter_2008_df.to_csv('better_2008_df.csv')\n"
```

This process was done for all years from 2008-2023.

Will concat all 'better' DF in data_cleaning_notebook_4

In [2]:
```
1 """
2 all_2010_stats_df = pd.read_csv('~/Documents/Flatiron/Project_5_/data/
3 """
```

In [3]:

```
"""
all_2010_stats_df.info()
"""
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 474758 entries, 0 to 474757
Data columns (total 92 columns):
 #   Column                          Non-Null Count   Dtype  
---  ------                          --------------   -----  
 0   pitch_type                      471786 non-null  object 
 1   game_date                       474758 non-null  object 
 2   release_speed                   471783 non-null  float64
 3   release_pos_x                   471770 non-null  float64
 4   release_pos_z                   471770 non-null  float64
 5   player_name                     474758 non-null  object 
 6   batter                          474758 non-null  int64  
 7   pitcher                         474758 non-null  int64  
 8   events                          124976 non-null  object 
 9   description                     474758 non-null  object 
 10  spin_dir                        0 non-null       float64
 11  spin_rate_deprecated            0 non-null       float64
 12  break_angle_deprecated          0 non-null       float64
 13  break_length_deprecated         0 non-null       float64
 14  zone                            471783 non-null  float64
 15  des                             474757 non-null  object 
 16  game_type                       474758 non-null  object 
 17  stand                           474758 non-null  object 
 18  p_throws                        474758 non-null  object 
 19  home_team                       474758 non-null  object 
 20  away_team                       474758 non-null  object 
 21  type                            474758 non-null  object 
 22  hit_location                    110347 non-null  float64
 23  bb_type                         91270 non-null   object 
 24  balls                           474758 non-null  int64  
 25  strikes                         474758 non-null  int64  
 26  game_year                       474758 non-null  int64  
 27  pfx_x                           471770 non-null  float64
 28  pfx_z                           471770 non-null  float64
 29  plate_x                         471783 non-null  float64
 30  plate_z                         471783 non-null  float64
 31  on_3b                           44276 non-null   float64
 32  on_2b                           87596 non-null   float64
 33  on_1b                           140420 non-null  float64
 34  outs_when_up                    474758 non-null  int64  
 35  inning                          474758 non-null  int64  
 36  inning_topbot                   474758 non-null  object 
 37  hc_x                            84545 non-null   float64
 38  hc_y                            84545 non-null   float64
 39  tfs_deprecated                  0 non-null       float64
 40  tfs_zulu_deprecated             0 non-null       float64
 41  fielder_2                       474758 non-null  int64  
 42  umpire                          0 non-null       float64
 43  sv_id                           471806 non-null  object 
 44  vx0                             471770 non-null  float64
 45  vy0                             471770 non-null  float64
 46  vz0                             471770 non-null  float64
 47  ax                              471783 non-null  float64
 48  ay                              471783 non-null  float64
 49  az                              471783 non-null  float64
 50  sz_top                          471783 non-null  float64
 51  sz_bot                          471783 non-null  float64
```

```
 52  hit_distance_sc                  0 non-null       float64
 53  launch_speed                     0 non-null       float64
 54  launch_angle                     0 non-null       float64
 55  effective_speed                  0 non-null       float64
 56  release_spin_rate                0 non-null       float64
 57  release_extension                0 non-null       float64
 58  game_pk                          474758 non-null  int64
 59  pitcher.1                        474758 non-null  int64
 60  fielder_2.1                      474758 non-null  int64
 61  fielder_3                        474758 non-null  int64
 62  fielder_4                        474758 non-null  int64
 63  fielder_5                        474758 non-null  int64
 64  fielder_6                        474758 non-null  int64
 65  fielder_7                        474758 non-null  int64
 66  fielder_8                        474758 non-null  int64
 67  fielder_9                        474758 non-null  int64
 68  release_pos_y                    471770 non-null  float64
 69  estimated_ba_using_speedangle    0 non-null       float64
 70  estimated_woba_using_speedangle  0 non-null       float64
 71  woba_value                       124977 non-null  float64
 72  woba_denom                       0 non-null       float64
 73  babip_value                      124977 non-null  float64
 74  iso_value                        124977 non-null  float64
 75  launch_speed_angle               0 non-null       float64
 76  at_bat_number                    474758 non-null  int64
 77  pitch_number                     474758 non-null  int64
 78  pitch_name                       471786 non-null  object
 79  home_score                       474758 non-null  int64
 80  away_score                       474758 non-null  int64
 81  bat_score                        474758 non-null  int64
 82  fld_score                        474758 non-null  int64
 83  post_away_score                  474758 non-null  int64
 84  post_home_score                  474758 non-null  int64
 85  post_bat_score                   474758 non-null  int64
 86  post_fld_score                   474758 non-null  int64
 87  if_fielding_alignment            0 non-null       float64
 88  of_fielding_alignment            0 non-null       float64
 89  spin_axis                        0 non-null       float64
 90  delta_home_win_exp               474758 non-null  float64
 91  delta_run_exp                    472124 non-null  float64
dtypes: float64(48), int64(28), object(16)
memory usage: 336.9+ MB
```

In [4]:

```
"""
all_2010_stats_df.drop(columns=['batter', 'events', 'description', 'zo
                                'des', 'game_type', 'stand', 'home_tea
                                'away_team', 'type', 'hit_location', '
                                'balls', 'strikes', 'pfx_x', 'spin_dir
                                'pfx_z', 'plate_x', 'plate_z', 'on_3b'
                                'on_2b', 'on_1b', 'outs_when_up', 'inn
                                'inning_topbot', 'hc_x', 'hc_y', 'fiel
                                'umpire', 'sv_id', 'hit_distance_sc',
                                 'sz_bot', 'launch_speed', 'launch_ang
                                'pitcher.1', 'fielder_2.1', 'fielder_3
                                'fielder_5', 'fielder_6', 'fielder_7',
                                'fielder_9', 'estimated_ba_using_speed
                                'estimated_woba_using_speedangle', 'ba
                                'launch_speed_angle', 'woba_value', 'w
                                'at_bat_number', 'pitch_number', 'home_
                                'bat_score', 'fld_score', 'post_home_s
                                'post_fld_score', 'post_away_score', '
                                'of_fielding_alignment', 'delta_home_w
                                'delta_run_exp', 'spin_rate_deprecated
                               'break_length_deprecated', 'tfs_deprecat
all_2010_stats_df.head()
"""
```

Out[4]:

	pitch_type	game_date	release_speed	release_pos_x	release_pos_z	player_name	pitche
0	SL	2010-10-03	82.8	1.82	6.45	Rhodes, Arthur	12112
1	SL	2010-10-03	81.1	1.76	6.46	Rhodes, Arthur	12112
2	FF	2010-10-03	90.3	1.94	6.31	Rhodes, Arthur	12112
3	SL	2010-10-03	80.4	1.71	6.55	Rhodes, Arthur	12112
4	SL	2010-10-03	80.7	1.76	6.52	Rhodes, Arthur	12112

5 rows × 21 columns

In [5]:
```
"""
all_2010_stats_df.info()
"""
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 474758 entries, 0 to 474757
Data columns (total 21 columns):
 #   Column             Non-Null Count   Dtype  
---  ------             --------------   -----  
 0   pitch_type         471786 non-null  object 
 1   game_date          474758 non-null  object 
 2   release_speed      471783 non-null  float64
 3   release_pos_x      471770 non-null  float64
 4   release_pos_z      471770 non-null  float64
 5   player_name        474758 non-null  object 
 6   pitcher            474758 non-null  int64  
 7   p_throws           474758 non-null  object 
 8   game_year          474758 non-null  int64  
 9   vx0                471770 non-null  float64
 10  vy0                471770 non-null  float64
 11  vz0                471770 non-null  float64
 12  ax                 471783 non-null  float64
 13  ay                 471783 non-null  float64
 14  az                 471783 non-null  float64
 15  effective_speed    0 non-null       float64
 16  release_spin_rate  0 non-null       float64
 17  release_extension  0 non-null       float64
 18  release_pos_y      471770 non-null  float64
 19  pitch_name         471786 non-null  object 
 20  spin_axis          0 non-null       float64
dtypes: float64(14), int64(2), object(5)
memory usage: 79.7+ MB
```

In [5]:
```
"""
all_2010_stats_df.drop(columns=['spin_axis', 'effective_speed',
                                'release_spin_rate', 'release_extension
all_2010_stats_df.head()
"""
```

Out[5]:

	pitch_type	game_date	release_speed	release_pos_x	release_pos_z	player_name	pitche
0	SL	2010-10-03	82.8	1.82	6.45	Rhodes, Arthur	12112
1	SL	2010-10-03	81.1	1.76	6.46	Rhodes, Arthur	12112
2	FF	2010-10-03	90.3	1.94	6.31	Rhodes, Arthur	12112
3	SL	2010-10-03	80.4	1.71	6.55	Rhodes, Arthur	12112
4	SL	2010-10-03	80.7	1.76	6.52	Rhodes, Arthur	12112

In [6]:
```
"""
all_2010_stats_df = all_2010_stats_df.dropna(axis=0)
"""
```

In [7]:
```
"""
all_2010_stats_df.reset_index(inplace=True)
"""
```

In [8]:
```
"""
all_2010_stats_df.drop('index', axis=1)
"""
```

Out[8]:

	pitch_type	game_date	release_speed	release_pos_x	release_pos_z	player_name
0	SL	2010-10-03	82.8	1.82	6.45	Rhodes, Arthur
1	SL	2010-10-03	81.1	1.76	6.46	Rhodes, Arthur
2	FF	2010-10-03	90.3	1.94	6.31	Rhodes, Arthur
3	SL	2010-10-03	80.4	1.71	6.55	Rhodes, Arthur
4	SL	2010-10-03	80.7	1.76	6.52	Rhodes, Arthur
...	...	...	...	...	...	...
471765	SI	2010-04-07	91.6	-1.45	6.52	Wainwright, Adam
471766	SI	2010-04-07	92.1	-1.29	6.58	Wainwright, Adam
471767	SI	2010-04-07	91.8	-1.33	6.57	Wainwright, Adam
471768	SI	2010-04-07	90.7	-1.28	6.65	Wainwright, Adam
471769	SI	2010-04-07	92.2	-1.20	6.58	Wainwright, Adam

471770 rows × 17 columns

In [10]:

```
"""
# Group by 'game_date' and 'pitcher' to calculate the total pitches
total_pitches = all_2010_stats_df.groupby(['game_date', 'pitcher', 'pla

# Group by 'game_date', 'pitcher', and 'pitch_type' to calculate the su
total_pitches_by_type = all_2010_stats_df.groupby(['game_date', 'pitch

# Calculate averages of the specified metrics for each pitch type, grou
avg_metrics = all_2010_stats_df.groupby(['game_date', 'pitcher', 'play
    'release_speed': 'mean',
    'release_pos_x': 'mean',
    'release_pos_z': 'mean',
    'vx0': 'mean',
    'vy0': 'mean',
    'vz0': 'mean',
    'ax': 'mean',
    'ay': 'mean',
    'az': 'mean',
    'release_pos_y': 'mean',
}).reset_index()

grouped_2010_df = total_pitches.merge(total_pitches_by_type, on=['game_
grouped_2010_df = grouped_2010_df.merge(avg_metrics, on=['game_date',

grouped_2010_df
"""
```

Out[10]:

	game_date	pitcher	player_name	total_pitches	pitch_type	count_by_pitch_type	rele
0	2010-04-04	120221	Park, Chan Ho	21	CH	2	
1	2010-04-04	120221	Park, Chan Ho	21	CU	1	
2	2010-04-04	120221	Park, Chan Ho	21	FF	3	
3	2010-04-04	120221	Park, Chan Ho	21	SI	8	
4	2010-04-04	120221	Park, Chan Ho	21	SL	7	
...	...	...	...	...	...	...	
30332	2010-10-03	502009	Latos, Mat	82	SI	20	
30333	2010-10-03	502009	Latos, Mat	82	SL	20	
30334	2010-10-03	519242	Sale, Chris	30	CH	1	
30335	2010-10-03	519242	Sale, Chris	30	SI	16	
30336	2010-10-03	519242	Sale, Chris	30	SL	13	

30337 rows × 16 columns

In [11]:

```
"""
grouped_2010_df['game_date'] = pd.to_datetime(grouped_2010_df['game_da
grouped_2010_df['season'] = grouped_2010_df['game_date'].dt.year

# Step 1: Season Total Pitches
season_total_pitches = grouped_2010_df.groupby(['pitcher', 'player_nam

# Step 2: Season Total by Pitch Type
season_total_by_pitch_type = grouped_2010_df.groupby(['pitcher', 'play

# Weighted Averages Calculation Setup
weighted_avg_columns = ['release_speed', 'release_pos_x', 'release_pos_
for col in weighted_avg_columns:
    grouped_2010_df[f'{col}_product'] = grouped_2010_df[col] * grouped_

weighted_avg_aggregations = {f'{col}_product': 'sum' for col in weight
weighted_avg_aggregations['count_by_pitch_type'] = 'sum'

# Aggregate for weighted averages
weighted_avg_df = grouped_2010_df.groupby(['pitcher', 'player_name', '

# Calculate weighted averages
for col in weighted_avg_columns:
    weighted_avg_df[f'{col}_weighted_avg'] = weighted_avg_df[f'{col}_p

# Cleanup
weighted_avg_df.drop(columns=[f'{col}_product' for col in weighted_avg_

# Merge season totals and weighted averages
final_2010_df = pd.merge(season_total_pitches, season_total_by_pitch_ty
final_2010_df = pd.merge(final_2010_df, weighted_avg_df, on=['pitcher'

final_2010_df.head()
"""
```

Out[11]:

	pitcher	player_name	season	season_total_pitches	pitch_type	season_total_count_by_pit
0	110683	Batista, Miguel	2010	4616	CH	
1	110683	Batista, Miguel	2010	4616	CU	
2	110683	Batista, Miguel	2010	4616	FF	
3	110683	Batista, Miguel	2010	4616	IN	
4	110683	Batista, Miguel	2010	4616	PO	

In [12]:
```
1 #final_2010_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1587 entries, 0 to 1586
Data columns (total 17 columns):
 #   Column                            Non-Null Count  Dtype  
---  ------                            --------------  -----  
 0   pitcher                           1587 non-null   int64  
 1   player_name                       1587 non-null   object 
 2   season                            1587 non-null   int32  
 3   season_total_pitches              1587 non-null   int64  
 4   pitch_type                        1587 non-null   object 
 5   season_total_count_by_pitch_type  1587 non-null   int64  
 6   count_by_pitch_type               1587 non-null   int64  
 7   release_speed_weighted_avg        1587 non-null   float64
 8   release_pos_x_weighted_avg        1587 non-null   float64
 9   release_pos_z_weighted_avg        1587 non-null   float64
 10  vx0_weighted_avg                  1587 non-null   float64
 11  vy0_weighted_avg                  1587 non-null   float64
 12  vz0_weighted_avg                  1587 non-null   float64
 13  ax_weighted_avg                   1587 non-null   float64
 14  ay_weighted_avg                   1587 non-null   float64
 15  az_weighted_avg                   1587 non-null   float64
 16  release_pos_y_weighted_avg        1587 non-null   float64
dtypes: float64(10), int32(1), int64(4), object(2)
memory usage: 204.7+ KB
```

In [13]:
```
1 #final_2010_df
```

Out[13]:

	pitcher	player_name	season	season_total_pitches	pitch_type	season_total_count_by
0	110683	Batista, Miguel	2010	4616	CH	
1	110683	Batista, Miguel	2010	4616	CU	
2	110683	Batista, Miguel	2010	4616	FF	
3	110683	Batista, Miguel	2010	4616	IN	
4	110683	Batista, Miguel	2010	4616	PO	
...	...	...	...	...	...	
1582	544931	Strasburg, Stephen	2010	4444	SL	
1583	545404	Beachy, Brandon	2010	1020	CH	
1584	545404	Beachy, Brandon	2010	1020	CU	
1585	545404	Beachy, Brandon	2010	1020	FF	
1586	545404	Beachy, Brandon	2010	1020	IN	

1587 rows × 17 columns

In [20]:

```
"""
final_2010_df['player_name'] = final_2010_df['player_name'].str.lower(
final_2010_df
"""
```

Out[20]:

	pitcher	player_name	season	season_total_pitches	pitch_type	season_total_count_by
0	110683	batista, miguel	2010	4616	CH	
1	110683	batista, miguel	2010	4616	CU	
2	110683	batista, miguel	2010	4616	FF	
3	110683	batista, miguel	2010	4616	IN	
4	110683	batista, miguel	2010	4616	PO	
...	...	...	...	...	...	
1582	544931	strasburg, stephen	2010	4444	SL	
1583	545404	beachy, brandon	2010	1020	CH	
1584	545404	beachy, brandon	2010	1020	CU	
1585	545404	beachy, brandon	2010	1020	FF	
1586	545404	beachy, brandon	2010	1020	IN	

1587 rows × 17 columns

In [21]:

```
1 #print(final_2010_df['player_name'].unique())
```

```
['batista, miguel' 'carpenter, chris' 'dessens, elmer' 'hampton, mike'
 'hawkins, latroy' 'hernandez, livan' 'lowe, derek' 'millwood, kevin'
 'moehler, brian' 'moyer, jamie' 'oliver, darren' 'park, chan ho'
 'pettitte, andy' 'reyes, dennys' 'rhodes, arthur' 'springer, russ'
 'suppan, jeff' 'wakefield, tim' 'wright, jamey' 'dempster, ryan'
 'ortiz, russ' 'wood, kerry' 'vazquez, javier' 'pavano, carl'
 'chen, bruce' 'halladay, roy' 'ortiz, ramon' 'schoeneweis, scott'
 'wolf, randy' 'figueroa, nelson' 'benson, kris' 'davis, doug'
 'marquis, jason' 'burnett, a.j.' 'lilly, ted' 'westbrook, jake'
 'lopez, rodrigo' 'penny, brad' 'franklin, ryan' 'weaver, jeff'
 'zito, barry' 'hudson, tim' 'padilla, vicente' 'meche, gil'
 'durbin, chad' 'downs, scott' 'santana, johan' 'arroyo, bronson'
 'benoit, joaquín' 'beckett, josh' 'belisle, matt' 'garland, jon'
 'buehrle, mark' 'sabathia, cc' 'vargas, claudio' 'sheets, ben'
 'dickey, r.a.' 'pineiro, joel' 'affeldt, jeremy' 'lohse, kyle'
 'cook, aaron' 'oswalt, roy' 'silva, carlos' 'ramirez, horacio'
 'moseley, dustin' 'zambrano, carlos' 'lackey, john' 'de la rosa, jorge'
 'chacin, gustavo' 'lewis, colby' 'sosa, jorge' 'myers, brett'
 'peavy, jake' 'harang, aaron' 'pérez, oliver' 'hendrickson, mark'
 'robertson, nate' 'tallet, brian' 'guthrie, jeremy' 'wellemeyer, todd'
 'kuo, hung-chih' 'capuano, chris' 'carrasco, d.j.' 'ledezma, wilfredo'
 'contreras, jose' 'wainwright, adam' 'bonser, boof' 'bonderman, jeremy'
 'greinke, zack' 'harden, rich' 'floyd, gavin' 'willis, dontrelle'
 'haren, dan' 'hill, shawn' 'jackson, edwin' 'maine, john'
 'santana, ervin' 'narveson, chris' 'correia, kevin' 'mitre, sergio'
 'gaudin, chad' 'simon, alfredo' 'blanton, joe' 'snell, ian'
 'germano, justin' 'maholm, paul' 'cain, matt' 'hamels, cole'
 'kazmir, scott' 'stauffer, tim' 'young, chris' 'danks, john'
 'thompson, brad' 'hernandez, roberto' 'francis, jeff' 'hernández, félix'
 'bush, dave' 'loe, kameron' 'verlander, justin' 'liriano, francisco'
 'tejeda, robinson' 'saunders, joe' 'jiménez, ubaldo' 'hammel, jason'
 'burres, brian' 'rodriguez, wandy' 'mendoza, luis' 'sánchez, aníbal'
 'davies, kyle' 'rowland-smith, ryan' 'duke, zach' 'olsen, scott'
 'niemann, jeff' 'karstens, jeff' 'laffey, aaron' 'feldman, scott'
 'nolasco, ricky' 'marshall, sean' 'chavez, jesse' 'eveland, dana'
 'litsch, jesse' 'detwiler, ross' 'cecil, brett' 'bannister, brian'
 'blackburn, nick' 'parra, manny' 'hill, rich' 'bergesen, brad'
 'shields, james' 'wells, randy' 'garcía, jaime' 'harrell, lucas'
 'volquez, edinson' 'vargas, jason' 'weaver, jered' 'wilson, c.j.'
 'medlen, kris' 'fister, doug' 'matusz, brian' 'galarraga, armando'
 'billingsley, chad' 'davis, wade' 'gallardo, yovani' 'marcum, shaun'
 'owings, micah' 'lester, jon' 'kendrick, kyle' 'gorzelanny, tom'
 'kennedy, ian' 'miller, andrew' 'leblanc, wade' 'scherzer, max'
 'huff, david' 'lincecum, tim' 'buchholz, clay' 'morrow, brandon'
 'richard, clayton' 'arrieta, jake' 'villanueva, carlos' 'hensley, clay'
 'ohlendorf, ross' 'price, david' 'sánchez, jonathan' 'cueto, johnny'
 'bergmann, jason' 'bailey, homer' 'harrison, matt' 'jurrjens, jair'
 'reyes, jo-jo' "o'sullivan, sean" 'happ, j.a.' 'luebke, cory'
 'sonnanstine, andy' 'volstad, chris' 'tomlin, josh' 'lannan, john'
 'slowey, kevin' 'humber, philip' 'hochevar, luke' 'pelfrey, mike'
 'romero, ricky' 'braden, dallas' 'gonzález, gio' 'hughes, phil'
 'hanson, tommy' 'estrada, marco' 'paulino, felipe' 'morales, franklin'
 'doubront, felix' 'nova, iván' 'ogando, alexi' 'chacín, jhoulys'
 'rogers, esmil' 'carrasco, carlos' 'anderson, brett' 'worley, vance'
 'ross, tyson' 'drabek, kyle' 'wood, travis' 'masterson, justin'
 'hellickson, jeremy' 'niese, jonathon' 'kershaw, clayton'
 'cashner, andrew' 'duensing, brian' 'hunter, tommy' 'stammen, craig'
```

```
'garza, matt' 'kuroda, hiroki' 'matsuzaka, daisuke' 'pineda, michael'
'tillman, chris' 'minor, mike' 'latos, mat' 'norris, bud'
'samardzija, jeff' 'leake, mike' 'cahill, trevor' 'holland, derek'
'bumgarner, madison' 'gee, dillon' 'porcello, rick' 'sale, chris'
'zimmermann, jordan' 'hudson, daniel' 'strasburg, stephen'
'beachy, brandon']
```

In [34]:

```
"""
def remove_accents(input_str):
    nfkd_form = unicodedata.normalize('NFKD', input_str)
    return "".join([c for c in nfkd_form if not unicodedata.combining(

def clean_name(name):
    name = name.lower()
    name = remove_accents(name)
    name = re.sub(r'[-.]', '', name)
    name = re.sub(r'\s+', ' ', name).strip()
    return name

final_2010_df['player_name'] = final_2010_df['player_name'].apply(clea
"""
```

Out[34]:

```
'\ndef remove_accents(input_str):\n    nfkd_form = unicodedata.normalize
(\'NFKD\', input_str)\n    return "".join([c for c in nfkd_form if not un
icodedata.combining(c)])\n\ndef clean_name(name):\n    name = name.lower
()\n    name = remove_accents(name)\n    name = re.sub(r\'[-.]\', \'\', n
ame)\n    name = re.sub(r\'\\s+\', \' \', name).strip()\n    return name
\n\nfinal_2010_df[\'player_name\'] = final_2010_df[\'player_name\'].apply
(clean_name)\n'
```

In [23]:
```
#final_2010_df
```

Out[23]:

	pitcher	player_name	season	season_total_pitches	pitch_type	season_total_count_by
0	110683	batista, miguel	2010	4616	CH	
1	110683	batista, miguel	2010	4616	CU	
2	110683	batista, miguel	2010	4616	FF	
3	110683	batista, miguel	2010	4616	IN	
4	110683	batista, miguel	2010	4616	PO	
...	...	...	...	...	...	
1582	544931	strasburg, stephen	2010	4444	SL	
1583	545404	beachy, brandon	2010	1020	CH	
1584	545404	beachy, brandon	2010	1020	CU	
1585	545404	beachy, brandon	2010	1020	FF	
1586	545404	beachy, brandon	2010	1020	IN	

1587 rows × 17 columns

In [29]:
```
"""
# Convert 'player_name' from "last name, first name" to "first name la
final_2010_df['Name'] = final_2010_df['player_name'].apply(lambda x: '
"""
```

In [30]:
```
#final_2010_df
```

Out[30]:

	pitcher	player_name	season	season_total_pitches	pitch_type	season_total_count_by
0	110683	batista, miguel	2010	4616	CH	
1	110683	batista, miguel	2010	4616	CU	
2	110683	batista, miguel	2010	4616	FF	
3	110683	batista, miguel	2010	4616	IN	
4	110683	batista, miguel	2010	4616	PO	
...	...	...	...	...	...	
1582	544931	strasburg, stephen	2010	4444	SL	
1583	545404	beachy, brandon	2010	1020	CH	
1584	545404	beachy, brandon	2010	1020	CU	
1585	545404	beachy, brandon	2010	1020	FF	
1586	545404	beachy, brandon	2010	1020	IN	

1587 rows × 18 columns

In [35]:
```
"""
# Now, you can perform the merge using 'Name' and 'season' as the keys
better_2010_df = pd.merge(final_2010_df,
                          cleaning_filtered_df[['Name', 'season', 'Age
                          on=['Name', 'season'],
                          how='left')
"""
```

In [36]:
```
1 #better_2010_df
```

Out[36]:

	pitcher	player_name	season	season_total_pitches	pitch_type	season_total_count_by
0	110683	batista, miguel	2010	4616	CH	
1	110683	batista, miguel	2010	4616	CU	
2	110683	batista, miguel	2010	4616	FF	
3	110683	batista, miguel	2010	4616	IN	
4	110683	batista, miguel	2010	4616	PO	
...	...	...	...	...	...	
1588	544931	strasburg, stephen	2010	4444	SL	
1589	545404	beachy, brandon	2010	1020	CH	
1590	545404	beachy, brandon	2010	1020	CU	
1591	545404	beachy, brandon	2010	1020	FF	
1592	545404	beachy, brandon	2010	1020	IN	

1593 rows × 19 columns

In [37]:
```
1 #better_2010_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1593 entries, 0 to 1592
Data columns (total 19 columns):
 #   Column                            Non-Null Count  Dtype  
---  ------                            --------------  -----  
 0   pitcher                           1593 non-null   int64  
 1   player_name                       1593 non-null   object 
 2   season                            1593 non-null   int32  
 3   season_total_pitches              1593 non-null   int64  
 4   pitch_type                        1593 non-null   object 
 5   season_total_count_by_pitch_type  1593 non-null   int64  
 6   count_by_pitch_type               1593 non-null   int64  
 7   release_speed_weighted_avg        1593 non-null   float64
 8   release_pos_x_weighted_avg        1593 non-null   float64
 9   release_pos_z_weighted_avg        1593 non-null   float64
 10  vx0_weighted_avg                  1593 non-null   float64
 11  vy0_weighted_avg                  1593 non-null   float64
 12  vz0_weighted_avg                  1593 non-null   float64
 13  ax_weighted_avg                   1593 non-null   float64
 14  ay_weighted_avg                   1593 non-null   float64
 15  az_weighted_avg                   1593 non-null   float64
 16  release_pos_y_weighted_avg        1593 non-null   float64
 17  Name                              1593 non-null   object 
 18  Age                               1325 non-null   float64
dtypes: float64(11), int32(1), int64(4), object(3)
memory usage: 230.4+ KB
```

In [39]:
```
"""
# Filter the DataFrame to show only rows where 'Age' is NaN
nan_age_rows = better_2010_df[better_2010_df['Age'].isna()]
nan_age_rows
"""
```

Out[39]:

	pitcher	player_name	season	season_total_pitches	pitch_type	season_total_count_by
14	113296	dessens, elmer	2010	2892	CH	
15	113296	dessens, elmer	2010	2892	CU	
16	113296	dessens, elmer	2010	2892	FF	
17	113296	dessens, elmer	2010	2892	IN	
18	113296	dessens, elmer	2010	2892	PO	
...	...	...	...	...	...	
1490	501381	pineda, michael	2010	40	FF	
1491	501381	pineda, michael	2010	40	SL	
1567	519242	sale, chris	2010	966	CH	
1568	519242	sale, chris	2010	966	SI	
1569	519242	sale, chris	2010	966	SL	

268 rows × 19 columns

In [41]:
```
#better_2010_df.to_csv('data/better_2010_df.csv')
```

2009 DF

In [42]:
```
#all_2009_stats_df = pd.read_csv('~/Documents/Flatiron/Project_5_/data/
```

In [43]:

```
"""
all_2009_stats_df.drop(columns=['batter', 'events', 'description', 'zo
                                'des', 'game_type', 'stand', 'home_tea
                                'away_team', 'type', 'hit_location', '
                                'balls', 'strikes', 'pfx_x', 'spin_dir
                                'pfx_z', 'plate_x', 'plate_z', 'on_3b'
                                'on_2b', 'on_1b', 'outs_when_up', 'inn
                                'inning_topbot', 'hc_x', 'hc_y', 'fiel
                                'umpire', 'sv_id', 'hit_distance_sc',
                                 'sz_bot', 'launch_speed', 'launch_ang
                                'pitcher.1', 'fielder_2.1', 'fielder_3
                                'fielder_5', 'fielder_6', 'fielder_7',
                                'fielder_9', 'estimated_ba_using_speed
                                'estimated_woba_using_speedangle', 'ba
                                'launch_speed_angle', 'woba_value', 'w
                                'at_bat_number', 'pitch_number', 'home_
                                'bat_score', 'fld_score', 'post_home_s
                                'post_fld_score', 'post_away_score', '
                                'of_fielding_alignment', 'delta_home_w
                                'delta_run_exp', 'spin_rate_deprecated
                               'break_length_deprecated', 'tfs_deprecat
all_2009_stats_df.head()
"""
```

Out[43]:

	pitch_type	game_date	release_speed	release_pos_x	release_pos_z	player_name	pitche
0	SL	2009-09-22	80.1	1.67	6.67	Rhodes, Arthur	12112
1	SL	2009-09-22	80.7	1.57	6.62	Rhodes, Arthur	12112
2	FF	2009-09-22	90.9	1.71	6.57	Rhodes, Arthur	12112
3	SL	2009-09-22	80.7	1.73	6.67	Rhodes, Arthur	12112
4	SL	2009-09-22	82.5	1.60	6.67	Rhodes, Arthur	12112

5 rows × 21 columns

In [44]:
```
#all_2009_stats_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 448937 entries, 0 to 448936
Data columns (total 21 columns):
 #   Column             Non-Null Count   Dtype  
---  ------             --------------   -----  
 0   pitch_type         443096 non-null  object 
 1   game_date          448937 non-null  object 
 2   release_speed      443089 non-null  float64
 3   release_pos_x      443084 non-null  float64
 4   release_pos_z      443084 non-null  float64
 5   player_name        448937 non-null  object 
 6   pitcher            448937 non-null  int64  
 7   p_throws           448937 non-null  object 
 8   game_year          448937 non-null  int64  
 9   vx0                443084 non-null  float64
 10  vy0                443084 non-null  float64
 11  vz0                443084 non-null  float64
 12  ax                 443089 non-null  float64
 13  ay                 443089 non-null  float64
 14  az                 443089 non-null  float64
 15  effective_speed    0 non-null       float64
 16  release_spin_rate  0 non-null       float64
 17  release_extension  0 non-null       float64
 18  release_pos_y      443084 non-null  float64
 19  pitch_name         443096 non-null  object 
 20  spin_axis          0 non-null       float64
dtypes: float64(14), int64(2), object(5)
memory usage: 75.4+ MB
```

In [45]:
```
"""
all_2009_stats_df.drop(columns=['spin_axis', 'effective_speed',
                                'release_spin_rate', 'release_extension
all_2009_stats_df.head()
"""
```

Out[45]:

	pitch_type	game_date	release_speed	release_pos_x	release_pos_z	player_name	pitche
0	SL	2009-09-22	80.1	1.67	6.67	Rhodes, Arthur	12112
1	SL	2009-09-22	80.7	1.57	6.62	Rhodes, Arthur	12112
2	FF	2009-09-22	90.9	1.71	6.57	Rhodes, Arthur	12112
3	SL	2009-09-22	80.7	1.73	6.67	Rhodes, Arthur	12112
4	SL	2009-09-22	82.5	1.60	6.67	Rhodes, Arthur	12112

In [46]:
```
#all_2009_stats_df = all_2009_stats_df.dropna(axis=0)
```

In [49]:
```
#all_2009_stats_df.reset_index(inplace=True)
```

In [50]:
```
#all_2009_stats_df.drop('index', axis=1)
```

Out[50]:

	pitch_type	game_date	release_speed	release_pos_x	release_pos_z	player_name
0	SL	2009-09-22	80.1	1.67	6.67	Rhodes, Arthur
1	SL	2009-09-22	80.7	1.57	6.62	Rhodes, Arthur
2	FF	2009-09-22	90.9	1.71	6.57	Rhodes, Arthur
3	SL	2009-09-22	80.7	1.73	6.67	Rhodes, Arthur
4	SL	2009-09-22	82.5	1.60	6.67	Rhodes, Arthur
...	...	...	...	...	...	...
443079	SI	2009-04-06	94.1	-0.92	6.72	Wainwright, Adam
443080	FC	2009-04-06	86.2	-1.13	6.61	Wainwright, Adam
443081	SI	2009-04-06	91.6	-1.13	6.68	Wainwright, Adam
443082	SI	2009-04-06	94.0	-1.05	6.44	Wainwright, Adam
443083	SI	2009-04-06	92.2	-1.31	6.41	Wainwright, Adam

443084 rows × 17 columns

In [51]:

```
"""
# Group by 'game_date' and 'pitcher' to calculate the total pitches
total_pitches = all_2009_stats_df.groupby(['game_date', 'pitcher', 'pla

# Group by 'game_date', 'pitcher', and 'pitch_type' to calculate the su
total_pitches_by_type = all_2009_stats_df.groupby(['game_date', 'pitch

# Calculate averages of the specified metrics for each pitch type, grou
avg_metrics = all_2009_stats_df.groupby(['game_date', 'pitcher', 'play
    'release_speed': 'mean',
    'release_pos_x': 'mean',
    'release_pos_z': 'mean',
    'vx0': 'mean',
    'vy0': 'mean',
    'vz0': 'mean',
    'ax': 'mean',
    'ay': 'mean',
    'az': 'mean',
    'release_pos_y': 'mean',
}).reset_index()

grouped_2009_df = total_pitches.merge(total_pitches_by_type, on=['game_
grouped_2009_df = grouped_2009_df.merge(avg_metrics, on=['game_date',

grouped_2009_df
"""
```

Out[51]:

	game_date	pitcher	player_name	total_pitches	pitch_type	count_by_pitch_type	rele
0	2009-04-05	113961	Eyre, Scott	9	FF	6	
1	2009-04-05	113961	Eyre, Scott	9	SL	3	
2	2009-04-05	117955	Lowe, Derek	97	FC	1	
3	2009-04-05	117955	Lowe, Derek	97	SI	59	
4	2009-04-05	117955	Lowe, Derek	97	SL	37	
...	...	...	...	...	...	...	
28445	2009-10-06	435261	Miner, Zach	23	SL	6	
28446	2009-10-06	519144	Porcello, Rick	92	CH	7	
28447	2009-10-06	519144	Porcello, Rick	92	FF	38	
28448	2009-10-06	519144	Porcello, Rick	92	SI	40	
28449	2009-10-06	519144	Porcello, Rick	92	SL	7	

28450 rows × 16 columns

In [52]:

```
"""
grouped_2009_df['game_date'] = pd.to_datetime(grouped_2009_df['game_da
grouped_2009_df['season'] = grouped_2009_df['game_date'].dt.year

# Step 1: Season Total Pitches
season_total_pitches = grouped_2009_df.groupby(['pitcher', 'player_nam

# Step 2: Season Total by Pitch Type
season_total_by_pitch_type = grouped_2009_df.groupby(['pitcher', 'play

# Weighted Averages Calculation Setup
weighted_avg_columns = ['release_speed', 'release_pos_x', 'release_pos_
for col in weighted_avg_columns:
    grouped_2009_df[f'{col}_product'] = grouped_2009_df[col] * grouped_

weighted_avg_aggregations = {f'{col}_product': 'sum' for col in weight
weighted_avg_aggregations['count_by_pitch_type'] = 'sum'

# Aggregate for weighted averages
weighted_avg_df = grouped_2009_df.groupby(['pitcher', 'player_name', '

# Calculate weighted averages
for col in weighted_avg_columns:
    weighted_avg_df[f'{col}_weighted_avg'] = weighted_avg_df[f'{col}_p

# Cleanup
weighted_avg_df.drop(columns=[f'{col}_product' for col in weighted_avg_

# Merge season totals and weighted averages
final_2009_df = pd.merge(season_total_pitches, season_total_by_pitch_ty
final_2009_df = pd.merge(final_2009_df, weighted_avg_df, on=['pitcher'

final_2009_df.head()
"""
```

Out[52]:

	pitcher	player_name	season	season_total_pitches	pitch_type	season_total_count_by_pit
0	110683	Batista, Miguel	2009	4267	CH	
1	110683	Batista, Miguel	2009	4267	CU	
2	110683	Batista, Miguel	2009	4267	FF	
3	110683	Batista, Miguel	2009	4267	IN	
4	110683	Batista, Miguel	2009	4267	SI	

In [53]:
```
1 #final_2009_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1584 entries, 0 to 1583
Data columns (total 17 columns):
 #   Column                            Non-Null Count  Dtype  
---  ------                            --------------  -----  
 0   pitcher                           1584 non-null   int64  
 1   player_name                       1584 non-null   object 
 2   season                            1584 non-null   int32  
 3   season_total_pitches              1584 non-null   int64  
 4   pitch_type                        1584 non-null   object 
 5   season_total_count_by_pitch_type  1584 non-null   int64  
 6   count_by_pitch_type               1584 non-null   int64  
 7   release_speed_weighted_avg        1584 non-null   float64
 8   release_pos_x_weighted_avg        1584 non-null   float64
 9   release_pos_z_weighted_avg        1584 non-null   float64
 10  vx0_weighted_avg                  1584 non-null   float64
 11  vy0_weighted_avg                  1584 non-null   float64
 12  vz0_weighted_avg                  1584 non-null   float64
 13  ax_weighted_avg                   1584 non-null   float64
 14  ay_weighted_avg                   1584 non-null   float64
 15  az_weighted_avg                   1584 non-null   float64
 16  release_pos_y_weighted_avg        1584 non-null   float64
dtypes: float64(10), int32(1), int64(4), object(2)
memory usage: 204.3+ KB
```

In [54]:
```
1 #final_2009_df
```

Out[54]:

	pitcher	player_name	season	season_total_pitches	pitch_type	season_total_count_by
0	110683	Batista, Miguel	2009	4267	CH	
1	110683	Batista, Miguel	2009	4267	CU	
2	110683	Batista, Miguel	2009	4267	FF	
3	110683	Batista, Miguel	2009	4267	IN	
4	110683	Batista, Miguel	2009	4267	SI	
...	...	...	...	...	...	
1579	519455	Zimmermann, Jordan	2009	6204	SL	
1580	543339	Hudson, Daniel	2009	1144	CH	
1581	543339	Hudson, Daniel	2009	1144	CU	
1582	543339	Hudson, Daniel	2009	1144	FF	
1583	543339	Hudson, Daniel	2009	1144	SL	

1584 rows × 17 columns

In [55]:

```
"""
final_2009_df['player_name'] = final_2009_df['player_name'].str.lower(
final_2009_df
"""
```

Out[55]:

	pitcher	player_name	season	season_total_pitches	pitch_type	season_total_count_by
0	110683	batista, miguel	2009	4267	CH	
1	110683	batista, miguel	2009	4267	CU	
2	110683	batista, miguel	2009	4267	FF	
3	110683	batista, miguel	2009	4267	IN	
4	110683	batista, miguel	2009	4267	SI	
...	...	...	...	...	...	
1579	519455	zimmermann, jordan	2009	6204	SL	
1580	543339	hudson, daniel	2009	1144	CH	
1581	543339	hudson, daniel	2009	1144	CU	
1582	543339	hudson, daniel	2009	1144	FF	
1583	543339	hudson, daniel	2009	1144	SL	

1584 rows × 17 columns

In [56]:
```
1 #print(final_2010_df['player_name'].unique())
```

```
['batista, miguel' 'carpenter, chris' 'dessens, elmer' 'hampton, mike'
 'hawkins, latroy' 'hernandez, livan' 'lowe, derek' 'millwood, kevin'
 'moehler, brian' 'moyer, jamie' 'oliver, darren' 'park, chan ho'
 'pettitte, andy' 'reyes, dennys' 'rhodes, arthur' 'springer, russ'
 'suppan, jeff' 'wakefield, tim' 'wright, jamey' 'dempster, ryan'
 'ortiz, russ' 'wood, kerry' 'vazquez, javier' 'pavano, carl'
 'chen, bruce' 'halladay, roy' 'ortiz, ramon' 'schoeneweis, scott'
 'wolf, randy' 'figueroa, nelson' 'benson, kris' 'davis, doug'
 'marquis, jason' 'burnett, aj' 'lilly, ted' 'westbrook, jake'
 'lopez, rodrigo' 'penny, brad' 'franklin, ryan' 'weaver, jeff'
 'zito, barry' 'hudson, tim' 'padilla, vicente' 'meche, gil'
 'durbin, chad' 'downs, scott' 'santana, johan' 'arroyo, bronson'
 'benoit, joaquin' 'beckett, josh' 'belisle, matt' 'garland, jon'
 'buehrle, mark' 'sabathia, cc' 'vargas, claudio' 'sheets, ben'
 'dickey, ra' 'pineiro, joel' 'affeldt, jeremy' 'lohse, kyle'
 'cook, aaron' 'oswalt, roy' 'silva, carlos' 'ramirez, horacio'
 'moseley, dustin' 'zambrano, carlos' 'lackey, john' 'de la rosa, jorge'
 'chacin, gustavo' 'lewis, colby' 'sosa, jorge' 'myers, brett'
 'peavy, jake' 'harang, aaron' 'perez, oliver' 'hendrickson, mark'
 'robertson, nate' 'tallet, brian' 'guthrie, jeremy' 'wellemeyer, todd'
 'kuo, hungchih' 'capuano, chris' 'carrasco, dj' 'ledezma, wilfredo'
 'contreras, jose' 'wainwright, adam' 'bonser, boof' 'bonderman, jeremy'
 'greinke, zack' 'harden, rich' 'floyd, gavin' 'willis, dontrelle'
 'haren, dan' 'hill, shawn' 'jackson, edwin' 'maine, john'
 'santana, ervin' 'narveson, chris' 'correia, kevin' 'mitre, sergio'
 'gaudin, chad' 'simon, alfredo' 'blanton, joe' 'snell, ian'
 'germano, justin' 'maholm, paul' 'cain, matt' 'hamels, cole'
 'kazmir, scott' 'stauffer, tim' 'young, chris' 'danks, john'
 'thompson, brad' 'hernandez, roberto' 'francis, jeff' 'hernandez, felix'
 'bush, dave' 'loe, kameron' 'verlander, justin' 'liriano, francisco'
 'tejeda, robinson' 'saunders, joe' 'jimenez, ubaldo' 'hammel, jason'
 'burres, brian' 'rodriguez, wandy' 'mendoza, luis' 'sanchez, anibal'
 'davies, kyle' 'rowlandsmith, ryan' 'duke, zach' 'olsen, scott'
 'niemann, jeff' 'karstens, jeff' 'laffey, aaron' 'feldman, scott'
 'nolasco, ricky' 'marshall, sean' 'chavez, jesse' 'eveland, dana'
 'litsch, jesse' 'detwiler, ross' 'cecil, brett' 'bannister, brian'
 'blackburn, nick' 'parra, manny' 'hill, rich' 'bergesen, brad'
 'shields, james' 'wells, randy' 'garcia, jaime' 'harrell, lucas'
 'volquez, edinson' 'vargas, jason' 'weaver, jered' 'wilson, cj'
 'medlen, kris' 'fister, doug' 'matusz, brian' 'galarraga, armando'
 'billingsley, chad' 'davis, wade' 'gallardo, yovani' 'marcum, shaun'
 'owings, micah' 'lester, jon' 'kendrick, kyle' 'gorzelanny, tom'
 'kennedy, ian' 'miller, andrew' 'leblanc, wade' 'scherzer, max'
 'huff, david' 'lincecum, tim' 'buchholz, clay' 'morrow, brandon'
 'richard, clayton' 'arrieta, jake' 'villanueva, carlos' 'hensley, clay'
 'ohlendorf, ross' 'price, david' 'sanchez, jonathan' 'cueto, johnny'
 'bergmann, jason' 'bailey, homer' 'harrison, matt' 'jurrjens, jair'
 'reyes, jojo' "o'sullivan, sean" 'happ, ja' 'luebke, cory'
 'sonnanstine, andy' 'volstad, chris' 'tomlin, josh' 'lannan, john'
 'slowey, kevin' 'humber, philip' 'hochevar, luke' 'pelfrey, mike'
 'romero, ricky' 'braden, dallas' 'gonzalez, gio' 'hughes, phil'
 'hanson, tommy' 'estrada, marco' 'paulino, felipe' 'morales, franklin'
 'doubront, felix' 'nova, ivan' 'ogando, alexi' 'chacin, jhoulys'
 'rogers, esmil' 'carrasco, carlos' 'anderson, brett' 'worley, vance'
 'ross, tyson' 'drabek, kyle' 'wood, travis' 'masterson, justin'
 'hellickson, jeremy' 'niese, jonathon' 'kershaw, clayton'
 'cashner, andrew' 'duensing, brian' 'hunter, tommy' 'stammen, craig'
```

```
'garza, matt' 'kuroda, hiroki' 'matsuzaka, daisuke' 'pineda, michael'
'tillman, chris' 'minor, mike' 'latos, mat' 'norris, bud'
'samardzija, jeff' 'leake, mike' 'cahill, trevor' 'holland, derek'
'bumgarner, madison' 'gee, dillon' 'porcello, rick' 'sale, chris'
'zimmermann, jordan' 'hudson, daniel' 'strasburg, stephen'
'beachy, brandon']
```

In [58]:

```
"""
def remove_accents(input_str):
    nfkd_form = unicodedata.normalize('NFKD', input_str)
    return "".join([c for c in nfkd_form if not unicodedata.combining(

def clean_name(name):
    name = name.lower()
    name = remove_accents(name)
    name = re.sub(r'[-.]', '', name)
    name = re.sub(r'\s+', ' ', name).strip()
    return name

final_2009_df['player_name'] = final_2009_df['player_name'].apply(clea
"""
```

In [59]:

```
#final_2009_df
```

Out[59]:

	pitcher	player_name	season	season_total_pitches	pitch_type	season_total_count_by
0	110683	batista, miguel	2009	4267	CH	
1	110683	batista, miguel	2009	4267	CU	
2	110683	batista, miguel	2009	4267	FF	
3	110683	batista, miguel	2009	4267	IN	
4	110683	batista, miguel	2009	4267	SI	
...	...	...	...	...	...	
1579	519455	zimmermann, jordan	2009	6204	SL	
1580	543339	hudson, daniel	2009	1144	CH	
1581	543339	hudson, daniel	2009	1144	CU	
1582	543339	hudson, daniel	2009	1144	FF	
1583	543339	hudson, daniel	2009	1144	SL	

1584 rows × 17 columns

In [60]:
```
"""
# Convert 'player_name' from "last name, first name" to "first name la
final_2009_df['Name'] = final_2009_df['player_name'].apply(lambda x: '
"""
```

In [61]:
```
#final_2009_df
```

Out[61]:

	pitcher	player_name	season	season_total_pitches	pitch_type	season_total_count_by
0	110683	batista, miguel	2009	4267	CH	
1	110683	batista, miguel	2009	4267	CU	
2	110683	batista, miguel	2009	4267	FF	
3	110683	batista, miguel	2009	4267	IN	
4	110683	batista, miguel	2009	4267	SI	
...	...	...	...	...	...	
1579	519455	zimmermann, jordan	2009	6204	SL	
1580	543339	hudson, daniel	2009	1144	CH	
1581	543339	hudson, daniel	2009	1144	CU	
1582	543339	hudson, daniel	2009	1144	FF	
1583	543339	hudson, daniel	2009	1144	SL	

1584 rows × 18 columns

In [62]:
```
"""
# Now, you can perform the merge using 'Name' and 'season' as the keys
better_2009_df = pd.merge(final_2009_df,
                          cleaning_filtered_df[['Name', 'season', 'Age
                          on=['Name', 'season'],
                          how='left')
"""
```

In [63]:
```
1 #better_2009_df
```

Out[63]:

	pitcher	player_name	season	season_total_pitches	pitch_type	season_total_count_by
0	110683	batista, miguel	2009	4267	CH	
1	110683	batista, miguel	2009	4267	CU	
2	110683	batista, miguel	2009	4267	FF	
3	110683	batista, miguel	2009	4267	IN	
4	110683	batista, miguel	2009	4267	SI	
...	...	...	...	...	...	
1604	519455	zimmermann, jordan	2009	6204	SL	
1605	543339	hudson, daniel	2009	1144	CH	
1606	543339	hudson, daniel	2009	1144	CU	
1607	543339	hudson, daniel	2009	1144	FF	
1608	543339	hudson, daniel	2009	1144	SL	

1609 rows × 19 columns

In [64]:
```
1 #better_2009_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1609 entries, 0 to 1608
Data columns (total 19 columns):
 #   Column                            Non-Null Count  Dtype
---  ------                            --------------  -----
 0   pitcher                           1609 non-null   int64
 1   player_name                       1609 non-null   object
 2   season                            1609 non-null   int32
 3   season_total_pitches              1609 non-null   int64
 4   pitch_type                        1609 non-null   object
 5   season_total_count_by_pitch_type  1609 non-null   int64
 6   count_by_pitch_type               1609 non-null   int64
 7   release_speed_weighted_avg        1609 non-null   float64
 8   release_pos_x_weighted_avg        1609 non-null   float64
 9   release_pos_z_weighted_avg        1609 non-null   float64
 10  vx0_weighted_avg                  1609 non-null   float64
 11  vy0_weighted_avg                  1609 non-null   float64
 12  vz0_weighted_avg                  1609 non-null   float64
 13  ax_weighted_avg                   1609 non-null   float64
 14  ay_weighted_avg                   1609 non-null   float64
 15  az_weighted_avg                   1609 non-null   float64
 16  release_pos_y_weighted_avg        1609 non-null   float64
 17  Name                              1609 non-null   object
 18  Age                               1384 non-null   float64
dtypes: float64(11), int32(1), int64(4), object(3)
memory usage: 232.7+ KB
```

In [65]:
```
1 #better_2009_df.to_csv('data/better_2009_df.csv')
```

2011

In [84]:
```
1 #all_2011_stats_df = pd.read_csv('~/Documents/Flatiron/Project_5_/data/
```

In [85]:

```
"""
all_2011_stats_df.drop(columns=['batter', 'events', 'description', 'zo
                                'des', 'game_type', 'stand', 'home_tea
                                'away_team', 'type', 'hit_location', '
                                'balls', 'strikes', 'pfx_x', 'spin_dir
                                'pfx_z', 'plate_x', 'plate_z', 'on_3b'
                                'on_2b', 'on_1b', 'outs_when_up', 'inn
                                'inning_topbot', 'hc_x', 'hc_y', 'fiel
                                'umpire', 'sv_id', 'hit_distance_sc',
                                 'sz_bot', 'launch_speed', 'launch_ang
                                'pitcher.1', 'fielder_2.1', 'fielder_3
                                'fielder_5', 'fielder_6', 'fielder_7',
                                'fielder_9', 'estimated_ba_using_speed
                                'estimated_woba_using_speedangle', 'ba
                                'launch_speed_angle', 'woba_value', 'w
                                'at_bat_number', 'pitch_number', 'home_
                                'bat_score', 'fld_score', 'post_home_s
                                'post_fld_score', 'post_away_score', '
                                'of_fielding_alignment', 'delta_home_w
                                'delta_run_exp', 'spin_rate_deprecated
                               'break_length_deprecated', 'tfs_deprecat
all_2011_stats_df.head()
"""
```

Out[85]:

	pitch_type	game_date	release_speed	release_pos_x	release_pos_z	player_name	pitche
0	FF	2011-09-27	87.0	2.22	6.16	Rhodes, Arthur	12112
1	FF	2011-09-27	89.6	2.09	6.15	Rhodes, Arthur	12112
2	FF	2011-09-27	86.8	2.11	6.21	Rhodes, Arthur	12112
3	FF	2011-09-27	87.3	2.11	6.15	Rhodes, Arthur	12112
4	FF	2011-09-27	87.9	2.08	6.22	Rhodes, Arthur	12112

5 rows × 21 columns

In [86]:
```
"""
all_2011_stats_df.drop(columns=['spin_axis', 'effective_speed',
                                'release_spin_rate', 'release_extension
all_2011_stats_df.head()
"""
```

Out[86]:

	pitch_type	game_date	release_speed	release_pos_x	release_pos_z	player_name	pitche
0	FF	2011-09-27	87.0	2.22	6.16	Rhodes, Arthur	12112
1	FF	2011-09-27	89.6	2.09	6.15	Rhodes, Arthur	12112
2	FF	2011-09-27	86.8	2.11	6.21	Rhodes, Arthur	12112
3	FF	2011-09-27	87.3	2.11	6.15	Rhodes, Arthur	12112
4	FF	2011-09-27	87.9	2.08	6.22	Rhodes, Arthur	12112

In [87]:
```
#all_2011_stats_df = all_2011_stats_df.dropna(axis=0)
```

In [88]:
```
#all_2011_stats_df.reset_index(inplace=True)
```

In [89]:
```
1 #all_2011_stats_df.drop('index', axis=1)
```

Out[89]:

	pitch_type	game_date	release_speed	release_pos_x	release_pos_z	player_name
0	FF	2011-09-27	87.0	2.22	6.16	Rhodes, Arthur
1	FF	2011-09-27	89.6	2.09	6.15	Rhodes, Arthur
2	FF	2011-09-27	86.8	2.11	6.21	Rhodes, Arthur
3	FF	2011-09-27	87.3	2.11	6.15	Rhodes, Arthur
4	FF	2011-09-27	87.9	2.08	6.22	Rhodes, Arthur
...	...	...	...	...	...	...
463711	FF	2011-03-31	96.2	-2.03	6.66	Verlander, Justin
463712	FF	2011-03-31	94.9	-2.09	6.65	Verlander, Justin
463713	FF	2011-03-31	95.1	-2.03	6.69	Verlander, Justin
463714	FF	2011-03-31	94.8	-2.18	6.69	Verlander, Justin
463715	FF	2011-03-31	93.3	-2.33	6.79	Verlander, Justin

463716 rows × 17 columns

In [90]:

```
"""
# Group by 'game_date' and 'pitcher' to calculate the total pitches
total_pitches = all_2011_stats_df.groupby(['game_date', 'pitcher', 'pla

# Group by 'game_date', 'pitcher', and 'pitch_type' to calculate the su
total_pitches_by_type = all_2011_stats_df.groupby(['game_date', 'pitche

# Calculate averages of the specified metrics for each pitch type, grou
avg_metrics = all_2011_stats_df.groupby(['game_date', 'pitcher', 'playe
    'release_speed': 'mean',
    'release_pos_x': 'mean',
    'release_pos_z': 'mean',
    'vx0': 'mean',
    'vy0': 'mean',
    'vz0': 'mean',
    'ax': 'mean',
    'ay': 'mean',
    'az': 'mean',
    'release_pos_y': 'mean',
}).reset_index()

grouped_2011_df = total_pitches.merge(total_pitches_by_type, on=['game_
grouped_2011_df = grouped_2011_df.merge(avg_metrics, on=['game_date',

grouped_2011_df
"""
```

Out[90]:

	game_date	pitcher	player_name	total_pitches	pitch_type	count_by_pitch_type	rele
0	2011-03-31	110683	Batista, Miguel	24	CH	3	
1	2011-03-31	110683	Batista, Miguel	24	FC	1	
2	2011-03-31	110683	Batista, Miguel	24	FF	3	
3	2011-03-31	110683	Batista, Miguel	24	SI	16	
4	2011-03-31	110683	Batista, Miguel	24	SL	1	
...	...	...	...	...	...	...	
28227	2011-09-28	572070	Richards, Garrett	71	CH	7	
28228	2011-09-28	572070	Richards, Garrett	71	FC	8	
28229	2011-09-28	572070	Richards, Garrett	71	FF	21	
28230	2011-09-28	572070	Richards, Garrett	71	SI	15	
28231	2011-09-28	572070	Richards, Garrett	71	SL	20	

28232 rows × 16 columns

In [91]:

```
"""
grouped_2011_df['game_date'] = pd.to_datetime(grouped_2011_df['game_da
grouped_2011_df['season'] = grouped_2011_df['game_date'].dt.year

# Step 1: Season Total Pitches
season_total_pitches = grouped_2011_df.groupby(['pitcher', 'player_nam

# Step 2: Season Total by Pitch Type
season_total_by_pitch_type = grouped_2011_df.groupby(['pitcher', 'play

# Weighted Averages Calculation Setup
weighted_avg_columns = ['release_speed', 'release_pos_x', 'release_pos_
for col in weighted_avg_columns:
    grouped_2011_df[f'{col}_product'] = grouped_2011_df[col] * grouped_

weighted_avg_aggregations = {f'{col}_product': 'sum' for col in weight
weighted_avg_aggregations['count_by_pitch_type'] = 'sum'

# Aggregate for weighted averages
weighted_avg_df = grouped_2011_df.groupby(['pitcher', 'player_name', '

# Calculate weighted averages
for col in weighted_avg_columns:
    weighted_avg_df[f'{col}_weighted_avg'] = weighted_avg_df[f'{col}_p

# Cleanup
weighted_avg_df.drop(columns=[f'{col}_product' for col in weighted_avg_

# Merge season totals and weighted averages
final_2011_df = pd.merge(season_total_pitches, season_total_by_pitch_ty
final_2011_df = pd.merge(final_2011_df, weighted_avg_df, on=['pitcher'

final_2011_df.head()
"""
```

Out[91]:

	pitcher	player_name	season	season_total_pitches	pitch_type	season_total_count_by_pit
0	110683	Batista, Miguel	2011	3944	CH	
1	110683	Batista, Miguel	2011	3944	CU	
2	110683	Batista, Miguel	2011	3944	FC	
3	110683	Batista, Miguel	2011	3944	FF	
4	110683	Batista, Miguel	2011	3944	IN	

In [92]:
```
"""
final_2011_df['player_name'] = final_2011_df['player_name'].str.lower(
final_2011_df
"""
```

Out[92]:

	pitcher	player_name	season	season_total_pitches	pitch_type	season_total_count_by
0	110683	batista, miguel	2011	3944	CH	
1	110683	batista, miguel	2011	3944	CU	
2	110683	batista, miguel	2011	3944	FC	
3	110683	batista, miguel	2011	3944	FF	
4	110683	batista, miguel	2011	3944	IN	
...	...	...	...	...	...	
1593	572070	richards, garrett	2011	1268	FC	
1594	572070	richards, garrett	2011	1268	FF	
1595	572070	richards, garrett	2011	1268	PO	
1596	572070	richards, garrett	2011	1268	SI	
1597	572070	richards, garrett	2011	1268	SL	

1598 rows × 17 columns

In [93]:
```
1 #print(final_2011_df['player_name'].unique())
```

```
['batista, miguel' 'carpenter, chris' 'colon, bartolo' 'hawkins, latroy'
 'hernandez, livan' 'isringhausen, jason' 'lowe, derek' 'millwood, kevin'
 'oliver, darren' 'reyes, dennys' 'rhodes, arthur' 'tomko, brett'
 'wakefield, tim' 'wright, jamey' 'dempster, ryan' 'wood, kerry'
 'vazquez, javier' 'pavano, carl' 'chen, bruce' 'halladay, roy'
 'ortiz, ramon' 'wolf, randy' 'figueroa, nelson' 'nathan, joe'
 'davis, doug' 'marquis, jason' 'burnett, a.j.' 'lilly, ted'
 'westbrook, jake' 'lopez, rodrigo' 'penny, brad' 'franklin, ryan'
 'zito, barry' 'hudson, tim' 'padilla, vicente' 'durbin, chad'
 'downs, scott' 'arroyo, bronson' 'benoit, joaquín' 'beckett, josh'
 'belisle, matt' 'garland, jon' 'buehrle, mark' 'sabathia, cc'
 'vogelsong, ryan' 'dickey, r.a.' 'pineiro, joel' 'affeldt, jeremy'
 'lohse, kyle' 'cook, aaron' 'oswalt, roy' 'ramirez, horacio'
 'moseley, dustin' 'zambrano, carlos' 'lackey, john' 'de la rosa, jorge'
 'bedard, erik' 'lewis, colby' 'myers, brett' 'peavy, jake'
 'harang, aaron' 'hendrickson, mark' 'tallet, brian' 'guthrie, jeremy'
 'wang, chien-ming' 'williams, jerome' 'kuo, hung-chih' 'capuano, chris'
 'carrasco, d.j.' 'ledezma, wilfredo' 'contreras, jose' 'greinke, zack'
 'harden, rich' 'floyd, gavin' 'willis, dontrelle' 'haren, dan'
 'jackson, edwin' 'santana, ervin' 'narveson, chris' 'correia, kevin'
 'mitre, sergio' 'gaudin, chad' 'simon, alfredo' 'blanton, joe'
 'mcgowan, dustin' 'germano, justin' 'maholm, paul' 'cain, matt'
 'hamels, cole' 'kazmir, scott' 'stauffer, tim' 'young, chris'
 'danks, john' 'hernandez, roberto' 'francis, jeff' 'hernández, félix'
 'bush, dave' 'loe, kameron' 'verlander, justin' 'liriano, francisco'
 'tejeda, robinson' 'saunders, joe' 'jiménez, ubaldo' 'hammel, jason'
 'burres, brian' 'rodriguez, wandy' 'mendoza, luis' 'sánchez, aníbal'
 'davies, kyle' 'duke, zach' 'mccarthy, brandon' 'niemann, jeff'
 'karstens, jeff' 'laffey, aaron' 'feldman, scott' 'nolasco, ricky'
 'marshall, sean' 'stults, eric' 'chavez, jesse' 'eveland, dana'
 'litsch, jesse' 'detwiler, ross' 'kluber, corey' 'cecil, brett'
 'blackburn, nick' 'hill, rich' 'bergesen, brad' 'shields, james'
 'wells, randy' 'garcía, jaime' 'harrell, lucas' 'volquez, edinson'
 'vargas, jason' 'weaver, jered' 'wilson, c.j.' 'medlen, kris'
 'fister, doug' 'matusz, brian' 'galarraga, armando' 'billingsley, chad'
 'davis, wade' 'gallardo, yovani' 'marcum, shaun' 'owings, micah'
 'lester, jon' 'kendrick, kyle' 'gorzelanny, tom' 'kennedy, ian'
 'miller, andrew' 'leblanc, wade' 'scherzer, max' 'huff, david'
 'lincecum, tim' 'buchholz, clay' 'morrow, brandon' 'richard, clayton'
 'arrieta, jake' 'villanueva, carlos' 'hensley, clay' 'ohlendorf, ross'
 'price, david' 'sánchez, jonathan' 'noesí, hector' 'cueto, johnny'
 'bailey, homer' 'harrison, matt' 'jurrjens, jair' 'reyes, jo-jo'
"o'sullivan, sean" 'happ, j.a.' 'luebke, cory' 'sonnanstine, andy'
 'lynn, lance' 'volstad, chris' 'tomlin, josh' 'lannan, john'
 'slowey, kevin' 'humber, philip' 'hochevar, luke' 'pelfrey, mike'
 'romero, ricky' 'braden, dallas' 'gonzález, gio' 'hughes, phil'
 'swarzak, anthony' 'hanson, tommy' 'estrada, marco' 'paulino, felipe'
 'morales, franklin' 'doubront, felix' 'nova, iván' 'ogando, alexi'
 'chacín, jhoulys' 'rogers, esmil' 'carrasco, carlos' 'anderson, brett'
 'worley, vance' 'ross, tyson' 'drabek, kyle' 'wood, travis'
 'masterson, justin' 'hellickson, jeremy' 'niese, jonathon'
 'kershaw, clayton' 'cashner, andrew' 'duensing, brian' 'hunter, tommy'
 'miley, wade' 'stammen, craig' 'garza, matt' 'kuroda, hiroki'
 'matsuzaka, daisuke' 'pineda, michael' 'tillman, chris' 'minor, mike'
 'latos, mat' 'norris, bud' 'locke, jeff' 'mcallister, zach' 'cobb, alex'
 'samardzija, jeff' 'leake, mike' 'cahill, trevor' 'santiago, héctor'
 'holland, derek' 'peacock, brad' 'nicasio, juan' 'alvarez iii, henderso
```

```
n'
 'delgado, randall' 'bumgarner, madison' 'collmenter, josh' 'duffy, dann
y'
 'gee, dillon' 'moore, matt' 'parker, jarrod' 'pomeranz, drew'
 'porcello, rick' 'sale, chris' 'zimmermann, jordan' 'de la rosa, rubby'
 'teheran, julio' 'diamond, scott' 'chatwood, tyler' 'eovaldi, nathan'
 'hudson, daniel' 'lyles, jordan' 'milone, tommy' 'strasburg, stephen'
 'turner, jacob' 'beachy, brandon' 'fiers, mike' 'richards, garrett']
```

In [94]:

```
"""
def remove_accents(input_str):
    nfkd_form = unicodedata.normalize('NFKD', input_str)
    return "".join([c for c in nfkd_form if not unicodedata.combining(

def clean_name(name):
    name = name.lower()
    name = remove_accents(name)
    name = re.sub(r'[-.]', '', name)
    name = re.sub(r'\s+', ' ', name).strip()
    return name

final_2011_df['player_name'] = final_2011_df['player_name'].apply(clea
"""
```

In [95]:

```
"""
# Convert 'player_name' from "last name, first name" to "first name la
final_2011_df['Name'] = final_2011_df['player_name'].apply(lambda x: '
"""
```

In [96]:

```
"""
# Now, you can perform the merge using 'Name' and 'season' as the keys
better_2011_df = pd.merge(final_2011_df,
                          cleaning_filtered_df[['Name', 'season', 'Age
                          on=['Name', 'season'],
                          how='left')
"""
```

In [114]:
```
1 #better_2011_df
```

Out[114]:

	pitcher	player_name	season	season_total_pitches	pitch_type	season_total_count_by
0	110683	batista, miguel	2011	3944	CH	
1	110683	batista, miguel	2011	3944	CU	
2	110683	batista, miguel	2011	3944	FC	
3	110683	batista, miguel	2011	3944	FF	
4	110683	batista, miguel	2011	3944	IN	
...	...	...	...	...	...	
1612	572070	richards, garrett	2011	1268	FC	
1613	572070	richards, garrett	2011	1268	FF	
1614	572070	richards, garrett	2011	1268	PO	
1615	572070	richards, garrett	2011	1268	SI	
1616	572070	richards, garrett	2011	1268	SL	

1617 rows × 19 columns

In [115]:
```
1 #better_2011_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1617 entries, 0 to 1616
Data columns (total 19 columns):
 #   Column                            Non-Null Count  Dtype
---  ------                            --------------  -----
 0   pitcher                           1617 non-null   int64
 1   player_name                       1617 non-null   object
 2   season                            1617 non-null   int32
 3   season_total_pitches              1617 non-null   int64
 4   pitch_type                        1617 non-null   object
 5   season_total_count_by_pitch_type  1617 non-null   int64
 6   count_by_pitch_type               1617 non-null   int64
 7   release_speed_weighted_avg        1617 non-null   float64
 8   release_pos_x_weighted_avg        1617 non-null   float64
 9   release_pos_z_weighted_avg        1617 non-null   float64
 10  vx0_weighted_avg                  1617 non-null   float64
 11  vy0_weighted_avg                  1617 non-null   float64
 12  vz0_weighted_avg                  1617 non-null   float64
 13  ax_weighted_avg                   1617 non-null   float64
 14  ay_weighted_avg                   1617 non-null   float64
 15  az_weighted_avg                   1617 non-null   float64
 16  release_pos_y_weighted_avg        1617 non-null   float64
 17  Name                              1617 non-null   object
 18  Age                               1384 non-null   float64
dtypes: float64(11), int32(1), int64(4), object(3)
memory usage: 233.8+ KB
```

In [99]:
```
1 #better_2011_df.to_csv('data/better_2011_df.csv')
```

2012

In [100]:
```
1 #all_2012_stats_df = pd.read_csv('~/Documents/Flatiron/Project_5_/data/
```

In [101]:

```
"""
all_2012_stats_df.drop(columns=['batter', 'events', 'description', 'zo
                                'des', 'game_type', 'stand', 'home_tea
                                'away_team', 'type', 'hit_location', 'l
                                'balls', 'strikes', 'pfx_x', 'spin_dir
                                'pfx_z', 'plate_x', 'plate_z', 'on_3b'
                                'on_2b', 'on_1b', 'outs_when_up', 'inn
                                'inning_topbot', 'hc_x', 'hc_y', 'fiel
                                'umpire', 'sv_id', 'hit_distance_sc',
                                 'sz_bot', 'launch_speed', 'launch_ang
                                'pitcher.1', 'fielder_2.1', 'fielder_3
                                'fielder_5', 'fielder_6', 'fielder_7',
                                'fielder_9', 'estimated_ba_using_speed
                                'estimated_woba_using_speedangle', 'ba
                                'launch_speed_angle', 'woba_value', 'w
                                'at_bat_number', 'pitch_number', 'home_
                                'bat_score', 'fld_score', 'post_home_s
                                'post_fld_score', 'post_away_score', '
                                'of_fielding_alignment', 'delta_home_w
                                'delta_run_exp', 'spin_rate_deprecated
                               'break_length_deprecated', 'tfs_deprecat
all_2012_stats_df.head()
"""
```

Out[101]:

	pitch_type	game_date	release_speed	release_pos_x	release_pos_z	player_name	pitche
0	FF	2012-09-30	92.5	-2.02	6.22	Hawkins, LaTroy	11562
1	SL	2012-09-30	84.2	-2.23	6.31	Hawkins, LaTroy	11562
2	CU	2012-09-30	76.2	-1.94	6.59	Hawkins, LaTroy	11562
3	FF	2012-09-30	91.8	-1.97	6.56	Hawkins, LaTroy	11562
4	CU	2012-09-30	74.7	-1.83	6.68	Hawkins, LaTroy	11562

5 rows × 21 columns

In [102]:
```
"""
all_2012_stats_df.drop(columns=['spin_axis', 'effective_speed',
                                'release_spin_rate', 'release_extension
all_2012_stats_df.head()
"""
```

Out[102]:

	pitch_type	game_date	release_speed	release_pos_x	release_pos_z	player_name	pitche
0	FF	2012-09-30	92.5	-2.02	6.22	Hawkins, LaTroy	11562
1	SL	2012-09-30	84.2	-2.23	6.31	Hawkins, LaTroy	11562
2	CU	2012-09-30	76.2	-1.94	6.59	Hawkins, LaTroy	11562
3	FF	2012-09-30	91.8	-1.97	6.56	Hawkins, LaTroy	11562
4	CU	2012-09-30	74.7	-1.83	6.68	Hawkins, LaTroy	11562

In [104]:
```
#all_2012_stats_df = all_2012_stats_df.dropna(axis=0)
```

In [105]:
```
#all_2012_stats_df.reset_index(inplace=True)
```

In [106]:
```
1 #all_2012_stats_df.drop('index', axis=1)
```

Out[106]:

	pitch_type	game_date	release_speed	release_pos_x	release_pos_z	player_name
0	FF	2012-09-30	92.5	-2.02	6.22	Hawkins, LaTroy
1	SL	2012-09-30	84.2	-2.23	6.31	Hawkins, LaTroy
2	CU	2012-09-30	76.2	-1.94	6.59	Hawkins, LaTroy
3	FF	2012-09-30	91.8	-1.97	6.56	Hawkins, LaTroy
4	CU	2012-09-30	74.7	-1.83	6.68	Hawkins, LaTroy
...	...	...	...	...	...	...
463493	FC	2012-04-07	88.0	-1.35	6.47	Wainwright, Adam
463494	SI	2012-04-07	90.4	-1.58	6.17	Wainwright, Adam
463495	CU	2012-04-07	73.7	-1.26	6.46	Wainwright, Adam
463496	SI	2012-04-07	89.5	-1.54	6.32	Wainwright, Adam
463497	SI	2012-04-07	91.1	-1.37	6.37	Wainwright, Adam

463498 rows × 17 columns

In [107]:

```
"""
# Group by 'game_date' and 'pitcher' to calculate the total pitches
total_pitches = all_2012_stats_df.groupby(['game_date', 'pitcher', 'pla

# Group by 'game_date', 'pitcher', and 'pitch_type' to calculate the su
total_pitches_by_type = all_2012_stats_df.groupby(['game_date', 'pitch

# Calculate averages of the specified metrics for each pitch type, grou
avg_metrics = all_2012_stats_df.groupby(['game_date', 'pitcher', 'play
    'release_speed': 'mean',
    'release_pos_x': 'mean',
    'release_pos_z': 'mean',
    'vx0': 'mean',
    'vy0': 'mean',
    'vz0': 'mean',
    'ax': 'mean',
    'ay': 'mean',
    'az': 'mean',
    'release_pos_y': 'mean',
}).reset_index()

grouped_2012_df = total_pitches.merge(total_pitches_by_type, on=['game_
grouped_2012_df = grouped_2012_df.merge(avg_metrics, on=['game_date',

grouped_2012_df
"""
```

Out[107]:

	game_date	pitcher	player_name	total_pitches	pitch_type	count_by_pitch_type	rele
0	2012-03-28	460024	Hochevar, Luke	80	CH	8	
1	2012-03-28	460024	Hochevar, Luke	80	CU	5	
2	2012-03-28	460024	Hochevar, Luke	80	FC	18	
3	2012-03-28	460024	Hochevar, Luke	80	FF	18	
4	2012-03-28	460024	Hochevar, Luke	80	SI	22	
...	...	...	...	...	...	...	
29680	2012-10-03	571946	Miller, Shelby	72	CH	8	
29681	2012-10-03	571946	Miller, Shelby	72	CU	14	
29682	2012-10-03	571946	Miller, Shelby	72	FF	50	
29683	2012-10-03	592767	Smyly, Drew	19	KC	6	
29684	2012-10-03	592767	Smyly, Drew	19	SI	13	

29685 rows × 16 columns

In [108]:

```
"""
grouped_2012_df['game_date'] = pd.to_datetime(grouped_2012_df['game_da
grouped_2012_df['season'] = grouped_2012_df['game_date'].dt.year

# Step 1: Season Total Pitches
season_total_pitches = grouped_2012_df.groupby(['pitcher', 'player_nam

# Step 2: Season Total by Pitch Type
season_total_by_pitch_type = grouped_2012_df.groupby(['pitcher', 'play

# Weighted Averages Calculation Setup
weighted_avg_columns = ['release_speed', 'release_pos_x', 'release_pos_
for col in weighted_avg_columns:
    grouped_2012_df[f'{col}_product'] = grouped_2012_df[col] * grouped_

weighted_avg_aggregations = {f'{col}_product': 'sum' for col in weight
weighted_avg_aggregations['count_by_pitch_type'] = 'sum'

# Aggregate for weighted averages
weighted_avg_df = grouped_2012_df.groupby(['pitcher', 'player_name', '

# Calculate weighted averages
for col in weighted_avg_columns:
    weighted_avg_df[f'{col}_weighted_avg'] = weighted_avg_df[f'{col}_p

# Cleanup
weighted_avg_df.drop(columns=[f'{col}_product' for col in weighted_avg_

# Merge season totals and weighted averages
final_2012_df = pd.merge(season_total_pitches, season_total_by_pitch_ty
final_2012_df = pd.merge(final_2012_df, weighted_avg_df, on=['pitcher'

final_2012_df.head()
"""
```

Out[108]:

	pitcher	player_name	season	season_total_pitches	pitch_type	season_total_count_by_pit
0	110683	Batista, Miguel	2012	3991	CH	
1	110683	Batista, Miguel	2012	3991	CU	
2	110683	Batista, Miguel	2012	3991	FC	
3	110683	Batista, Miguel	2012	3991	FF	
4	110683	Batista, Miguel	2012	3991	IN	

In [109]:

```
"""
final_2012_df['player_name'] = final_2012_df['player_name'].str.lower(
final_2012_df
"""
```

Out[109]:

	pitcher	player_name	season	season_total_pitches	pitch_type	season_total_count_by
0	110683	batista, miguel	2012	3991	CH	
1	110683	batista, miguel	2012	3991	CU	
2	110683	batista, miguel	2012	3991	FC	
3	110683	batista, miguel	2012	3991	FF	
4	110683	batista, miguel	2012	3991	IN	
...	...	...	...	...	...	
1598	612672	chen, wei-yin	2012	15594	CU	
1599	612672	chen, wei-yin	2012	15594	FA	
1600	612672	chen, wei-yin	2012	15594	FF	
1601	612672	chen, wei-yin	2012	15594	SI	
1602	612672	chen, wei-yin	2012	15594	SL	

1603 rows × 17 columns

In [110]:
```
1 #print(final_2012_df['player_name'].unique())
```

```
['batista, miguel' 'carpenter, chris' 'colon, bartolo' 'hawkins, latroy'
 'hernandez, livan' 'isringhausen, jason' 'lowe, derek' 'millwood, kevin'
 'moyer, jamie' 'oliver, darren' 'pettitte, andy' 'suppan, jeff'
 'wright, jamey' 'dempster, ryan' 'wood, kerry' 'pavano, carl'
 'chen, bruce' 'halladay, roy' 'wolf, randy' 'nathan, joe'
 'marquis, jason' 'wells, kip' 'burnett, a.j.' 'lilly, ted'
 'westbrook, jake' 'lopez, rodrigo' 'glover, gary' 'penny, brad'
 'zito, barry' 'hudson, tim' 'padilla, vicente' 'durbin, chad'
 'downs, scott' 'santana, johan' 'arroyo, bronson' 'benoit, joaquín'
 'beckett, josh' 'belisle, matt' 'buehrle, mark' 'sabathia, cc'
 'sheets, ben' 'vogelsong, ryan' 'dickey, r.a.' 'affeldt, jeremy'
 'lohse, kyle' 'cook, aaron' 'oswalt, roy' 'moseley, dustin'
 'zambrano, carlos' 'de la rosa, jorge' 'bedard, erik' 'lewis, colby'
 'myers, brett' 'peavy, jake' 'harang, aaron' 'pérez, oliver'
 'guthrie, jeremy' 'wang, chien-ming' 'williams, jerome' 'capuano, chris'
 'carrasco, d.j.' 'contreras, jose' 'wainwright, adam' 'greinke, zack'
 'floyd, gavin' 'haren, dan' 'hill, shawn' 'jackson, edwin'
 'santana, ervin' 'narveson, chris' 'correia, kevin' 'gaudin, chad'
 'simon, alfredo' 'blanton, joe' 'germano, justin' 'maholm, paul'
 'cain, matt' 'hamels, cole' 'stauffer, tim' 'young, chris' 'danks, john'
 'hernandez, roberto' 'francis, jeff' 'hernández, félix' 'petit, yusmeir
o'
 'loe, kameron' 'verlander, justin' 'liriano, francisco' 'saunders, joe'
 'jiménez, ubaldo' 'hammel, jason' 'rodriguez, wandy' 'mendoza, luis'
 'sánchez, aníbal' 'duke, zach' 'mccarthy, brandon' 'niemann, jeff'
 'karstens, jeff' 'laffey, aaron' 'feldman, scott' 'nolasco, ricky'
 'marshall, sean' 'stults, eric' 'chavez, jesse' 'eveland, dana'
 'detwiler, ross' 'kluber, corey' 'cecil, brett' 'blackburn, nick'
 'parra, manny' 'hill, rich' 'bergesen, brad' 'shields, james'
 'wells, randy' 'garcía, jaime' 'harrell, lucas' 'volquez, edinson'
 'vargas, jason' 'weaver, jered' 'wilson, c.j.' 'medlen, kris'
 'fister, doug' 'matusz, brian' 'galarraga, armando' 'billingsley, chad'
 'davis, wade' 'gallardo, yovani' 'marcum, shaun' 'owings, micah'
 'lester, jon' 'kendrick, kyle' 'gorzelanny, tom' 'kennedy, ian'
 'miller, andrew' 'leblanc, wade' 'scherzer, max' 'huff, david'
 'lincecum, tim' 'buchholz, clay' 'morrow, brandon' 'richard, clayton'
 'arrieta, jake' 'villanueva, carlos' 'hensley, clay' 'ohlendorf, ross'
 'price, david' 'sánchez, jonathan' 'noesí, hector' 'griffin, a.j.'
 'cueto, johnny' 'bailey, homer' 'harrison, matt' 'jurrjens, jair'
 "o'sullivan, sean" 'happ, j.a.' 'luebke, cory' 'hefner, jeremy'
 'lynn, lance' 'volstad, chris' 'tomlin, josh' 'lannan, john'
 'humber, philip' 'hochevar, luke' 'pelfrey, mike' 'romero, ricky'
 'gonzález, gio' 'hughes, phil' 'swarzak, anthony' 'hanson, tommy'
 'estrada, marco' 'paulino, felipe' 'morales, franklin' 'doubront, felix'
 'nova, iván' 'ogando, alexi' 'chacín, jhoulys' 'rogers, esmil'
 'anderson, brett' 'worley, vance' 'ross, tyson' 'drabek, kyle'
 'wood, travis' 'masterson, justin' 'phelps, david' 'hellickson, jeremy'
 'niese, jonathon' 'kershaw, clayton' 'cashner, andrew' 'duensing, brian'
 'hunter, tommy' 'miley, wade' 'stammen, craig' 'garza, matt'
 'kuroda, hiroki' 'matsuzaka, daisuke' 'quintana, jose' 'tillman, chris'
 'minor, mike' 'latos, mat' 'norris, bud' 'archer, chris' 'locke, jeff'
 'mcallister, zach' 'cobb, alex' 'samardzija, jeff' 'leake, mike'
 'cahill, trevor' 'santiago, héctor' 'holland, derek' 'peralta, wily'
 'nicasio, juan' 'darvish, yu' 'alvarez iii, henderson' 'delgado, randal
l'
 'bumgarner, madison' 'collmenter, josh' 'duffy, danny' 'gee, dillon'
 'harvey, matt' 'moore, matt' 'oberholtzer, brett' 'parker, jarrod'
```

```
'pomeranz, drew' 'porcello, rick' 'sale, chris' 'zimmermann, jordan'
'de la rosa, rubby' 'pérez, martín' 'teheran, julio' 'diamond, scott'
'chatwood, tyler' 'eovaldi, nathan' 'hudson, daniel' 'koehler, tom'
'lyles, jordan' 'mchugh, collin' 'milone, tommy' 'odorizzi, jake'
'strasburg, stephen' 'bauer, trevor' 'turner, jacob' 'beachy, brandon'
'iwakuma, hisashi' 'corbin, patrick' 'fiers, mike' 'hutchison, drew'
'mikolas, miles' 'miller, shelby' 'richards, garrett' 'skaggs, tyler'
'keuchel, dallas' 'straily, dan' 'smyly, drew' 'bundy, dylan'
'chen, wei-yin']
```

In [111]:

```
"""
def remove_accents(input_str):
    nfkd_form = unicodedata.normalize('NFKD', input_str)
    return "".join([c for c in nfkd_form if not unicodedata.combining(

def clean_name(name):
    name = name.lower()
    name = remove_accents(name)
    name = re.sub(r'[-.]', '', name)
    name = re.sub(r'\s+', ' ', name).strip()
    return name

final_2012_df['player_name'] = final_2012_df['player_name'].apply(clea
"""
```

In [112]:

```
"""
# Convert 'player_name' from "last name, first name" to "first name la
final_2012_df['Name'] = final_2012_df['player_name'].apply(lambda x: '
"""
```

In [113]:

```
"""
# Now, you can perform the merge using 'Name' and 'season' as the keys
better_2012_df = pd.merge(final_2012_df,
                          cleaning_filtered_df[['Name', 'season', 'Age
                          on=['Name', 'season'],
                          how='left')
"""
```

In [116]:
```
1 #better_2012_df
```

Out[116]:

	pitcher	player_name	season	season_total_pitches	pitch_type	season_total_count_by
0	110683	batista, miguel	2012	3991	CH	
1	110683	batista, miguel	2012	3991	CH	
2	110683	batista, miguel	2012	3991	CU	
3	110683	batista, miguel	2012	3991	CU	
4	110683	batista, miguel	2012	3991	FC	
...	...	...	...	...	...	
1625	612672	chen, weiyin	2012	15594	CU	
1626	612672	chen, weiyin	2012	15594	FA	
1627	612672	chen, weiyin	2012	15594	FF	
1628	612672	chen, weiyin	2012	15594	SI	
1629	612672	chen, weiyin	2012	15594	SL	

1630 rows × 19 columns

In [117]:
```
1 #better_2012_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1630 entries, 0 to 1629
Data columns (total 19 columns):
 #   Column                            Non-Null Count  Dtype
---  ------                            --------------  -----
 0   pitcher                           1630 non-null   int64
 1   player_name                       1630 non-null   object
 2   season                            1630 non-null   int32
 3   season_total_pitches              1630 non-null   int64
 4   pitch_type                        1630 non-null   object
 5   season_total_count_by_pitch_type  1630 non-null   int64
 6   count_by_pitch_type               1630 non-null   int64
 7   release_speed_weighted_avg        1630 non-null   float64
 8   release_pos_x_weighted_avg        1630 non-null   float64
 9   release_pos_z_weighted_avg        1630 non-null   float64
 10  vx0_weighted_avg                  1630 non-null   float64
 11  vy0_weighted_avg                  1630 non-null   float64
 12  vz0_weighted_avg                  1630 non-null   float64
 13  ax_weighted_avg                   1630 non-null   float64
 14  ay_weighted_avg                   1630 non-null   float64
 15  az_weighted_avg                   1630 non-null   float64
 16  release_pos_y_weighted_avg        1630 non-null   float64
 17  Name                              1630 non-null   object
 18  Age                               1402 non-null   float64
dtypes: float64(11), int32(1), int64(4), object(3)
memory usage: 235.7+ KB
```

In [118]:
```
1 #better_2012_df.to_csv('data/better_2012_df.csv')
```

2013

In [119]:
```
1 #all_2013_stats_df = pd.read_csv('~/Documents/Flatiron/Project_5_/data/
```

In [120]:
```
"""all_2013_stats_df.drop(columns=['batter', 'events', 'description',
                                   'des', 'game_type', 'stand', 'home_tea
                                   'away_team', 'type', 'hit_location', '
                                   'balls', 'strikes', 'pfx_x', 'spin_dir
                                   'pfx_z', 'plate_x', 'plate_z', 'on_3b'
                                   'on_2b', 'on_1b', 'outs_when_up', 'inn
                                   'inning_topbot', 'hc_x', 'hc_y', 'fiel
                                   'umpire', 'sv_id', 'hit_distance_sc',
                                    'sz_bot', 'launch_speed', 'launch_ang
                                   'pitcher.1', 'fielder_2.1', 'fielder_3
                                   'fielder_5', 'fielder_6', 'fielder_7',
                                   'fielder_9', 'estimated_ba_using_speed
                                   'estimated_woba_using_speedangle', 'ba
                                   'launch_speed_angle', 'woba_value', 'w
                                   'at_bat_number', 'pitch_number', 'home_
                                   'bat_score', 'fld_score', 'post_home_s
                                   'post_fld_score', 'post_away_score', '
                                   'of_fielding_alignment', 'delta_home_w
                                   'delta_run_exp', 'spin_rate_deprecated
                                  'break_length_deprecated', 'tfs_deprecat
all_2013_stats_df.head()
"""
```

Out[120]:

	pitch_type	game_date	release_speed	release_pos_x	release_pos_z	player_name	pitche
0	SL	2013-09-28	87.5	-1.88	6.24	Hawkins, LaTroy	11562
1	FF	2013-09-28	93.7	-1.69	6.17	Hawkins, LaTroy	11562
2	FF	2013-09-28	93.4	-1.61	6.23	Hawkins, LaTroy	11562
3	SL	2013-09-28	88.0	-1.90	6.22	Hawkins, LaTroy	11562
4	FF	2013-09-28	93.3	-1.80	6.25	Hawkins, LaTroy	11562

5 rows × 21 columns

In [121]:
```
"""
all_2013_stats_df.drop(columns=['spin_axis', 'effective_speed',
                                'release_spin_rate', 'release_extension
"""
```

In [122]:
```
#all_2013_stats_df = all_2013_stats_df.dropna(axis=0)
```

In [123]:
```
#all_2013_stats_df.reset_index(inplace=True)
```

In [124]:
```
#all_2013_stats_df.drop('index', axis=1)
```

Out[124]:

	pitch_type	game_date	release_speed	release_pos_x	release_pos_z	player_name
0	SL	2013-09-28	87.5	-1.88	6.24	Hawkins, LaTroy
1	FF	2013-09-28	93.7	-1.69	6.17	Hawkins, LaTroy
2	FF	2013-09-28	93.4	-1.61	6.23	Hawkins, LaTroy
3	SL	2013-09-28	88.0	-1.90	6.22	Hawkins, LaTroy
4	FF	2013-09-28	93.3	-1.80	6.25	Hawkins, LaTroy
...	...	...	...	...	...	...
452986	KC	2013-05-30	81.9	2.38	5.43	Wood, Alex
452987	CH	2013-05-30	84.4	2.55	5.39	Wood, Alex
452988	CH	2013-05-30	85.4	2.30	5.45	Wood, Alex
452989	SI	2013-05-30	94.4	2.14	5.63	Wood, Alex
452990	SI	2013-05-30	94.8	2.31	5.52	Wood, Alex

452991 rows × 17 columns

In [125]:

```
"""
# Group by 'game_date' and 'pitcher' to calculate the total pitches
total_pitches = all_2013_stats_df.groupby(['game_date', 'pitcher', 'pla

# Group by 'game_date', 'pitcher', and 'pitch_type' to calculate the su
total_pitches_by_type = all_2013_stats_df.groupby(['game_date', 'pitch

# Calculate averages of the specified metrics for each pitch type, grou
avg_metrics = all_2013_stats_df.groupby(['game_date', 'pitcher', 'play
    'release_speed': 'mean',
    'release_pos_x': 'mean',
    'release_pos_z': 'mean',
    'vx0': 'mean',
    'vy0': 'mean',
    'vz0': 'mean',
    'ax': 'mean',
    'ay': 'mean',
    'az': 'mean',
    'release_pos_y': 'mean',
}).reset_index()

grouped_2013_df = total_pitches.merge(total_pitches_by_type, on=['game_
grouped_2013_df = grouped_2013_df.merge(avg_metrics, on=['game_date',

grouped_2013_df
"""
```

Out[125]:

	game_date	pitcher	player_name	total_pitches	pitch_type	count_by_pitch_type	rele
0	2013-03-31	117955	Lowe, Derek	10	SI	6	
1	2013-03-31	117955	Lowe, Derek	10	SL	4	
2	2013-03-31	407853	Bedard, Erik	38	CH	1	
3	2013-03-31	407853	Bedard, Erik	38	CU	4	
4	2013-03-31	407853	Bedard, Erik	38	FF	25	
...	...	...	...	...	...	...	
27722	2013-09-30	527048	Pérez, Martín	74	CH	22	
27723	2013-09-30	527048	Pérez, Martín	74	CU	4	
27724	2013-09-30	527048	Pérez, Martín	74	FF	22	
27725	2013-09-30	527048	Pérez, Martín	74	SI	13	
27726	2013-09-30	527048	Pérez, Martín	74	SL	13	

27727 rows × 16 columns

In [126]:

```
"""
grouped_2013_df['game_date'] = pd.to_datetime(grouped_2013_df['game_da
grouped_2013_df['season'] = grouped_2013_df['game_date'].dt.year

# Step 1: Season Total Pitches
season_total_pitches = grouped_2013_df.groupby(['pitcher', 'player_nam

# Step 2: Season Total by Pitch Type
season_total_by_pitch_type = grouped_2013_df.groupby(['pitcher', 'play

# Weighted Averages Calculation Setup
weighted_avg_columns = ['release_speed', 'release_pos_x', 'release_pos_
for col in weighted_avg_columns:
    grouped_2013_df[f'{col}_product'] = grouped_2013_df[col] * grouped_

weighted_avg_aggregations = {f'{col}_product': 'sum' for col in weight
weighted_avg_aggregations['count_by_pitch_type'] = 'sum'

# Aggregate for weighted averages
weighted_avg_df = grouped_2013_df.groupby(['pitcher', 'player_name', '

# Calculate weighted averages
for col in weighted_avg_columns:
    weighted_avg_df[f'{col}_weighted_avg'] = weighted_avg_df[f'{col}_p

# Cleanup
weighted_avg_df.drop(columns=[f'{col}_product' for col in weighted_avg_

# Merge season totals and weighted averages
final_2013_df = pd.merge(season_total_pitches, season_total_by_pitch_ty
final_2013_df = pd.merge(final_2013_df, weighted_avg_df, on=['pitcher'

final_2013_df.head()
"""
```

Out[126]:

	pitcher	player_name	season	season_total_pitches	pitch_type	season_total_count_by_pit
0	112526	Colon, Bartolo	2013	10929	CH	
1	112526	Colon, Bartolo	2013	10929	FF	
2	112526	Colon, Bartolo	2013	10929	SI	
3	112526	Colon, Bartolo	2013	10929	SL	
4	115629	Hawkins, LaTroy	2013	3189	CH	

In [127]:

```
"""
final_2013_df['player_name'] = final_2013_df['player_name'].str.lower(
final_2013_df
"""
```

Out[127]:

	pitcher	player_name	season	season_total_pitches	pitch_type	season_total_count_by
0	112526	colon, bartolo	2013	10929	CH	
1	112526	colon, bartolo	2013	10929	FF	
2	112526	colon, bartolo	2013	10929	SI	
3	112526	colon, bartolo	2013	10929	SL	
4	115629	hawkins, latroy	2013	3189	CH	
...	...	...	...	...	...	
1501	622072	wood, alex	2013	4803	FF	
1502	622072	wood, alex	2013	4803	IN	
1503	622072	wood, alex	2013	4803	KC	
1504	622072	wood, alex	2013	4803	PO	
1505	622072	wood, alex	2013	4803	SI	

1506 rows × 17 columns

In [128]:
```
1 #print(final_2013_df['player_name'].unique())
```

```
['colon, bartolo' 'hawkins, latroy' 'lowe, derek' 'oliver, darren'
 'pettitte, andy' 'wright, jamey' 'dempster, ryan' 'chen, bruce'
 'halladay, roy' 'ortiz, ramon' 'nathan, joe' 'marquis, jason'
 'burnett, a.j.' 'lilly, ted' 'westbrook, jake' 'zito, barry'
 'hudson, tim' 'durbin, chad' 'downs, scott' 'arroyo, bronson'
 'benoit, joaquín' 'beckett, josh' 'belisle, matt' 'garland, jon'
 'buehrle, mark' 'sabathia, cc' 'vogelsong, ryan' 'dickey, r.a.'
 'affeldt, jeremy' 'lohse, kyle' 'oswalt, roy' 'lackey, john'
 'de la rosa, jorge' 'bedard, erik' 'myers, brett' 'peavy, jake'
 'harang, aaron' 'pérez, oliver' 'guthrie, jeremy' 'wang, chien-ming'
 'williams, jerome' 'capuano, chris' 'contreras, jose' 'wainwright, adam'
 'bonderman, jeremy' 'greinke, zack' 'floyd, gavin' 'haren, dan'
 'jackson, edwin' 'maine, john' 'santana, ervin' 'narveson, chris'
 'correia, kevin' 'gaudin, chad' 'simon, alfredo' 'blanton, joe'
 'mcgowan, dustin' 'germano, justin' 'maholm, paul' 'cain, matt'
 'hamels, cole' 'kazmir, scott' 'stauffer, tim' 'danks, john'
 'hernandez, roberto' 'francis, jeff' 'hernández, félix' 'petit, yusmeir
o'
 'bush, dave' 'loe, kameron' 'verlander, justin' 'liriano, francisco'
 'saunders, joe' 'jiménez, ubaldo' 'hammel, jason' 'rodriguez, wandy'
 'mendoza, luis' 'sánchez, aníbal' 'duke, zach' 'mccarthy, brandon'
 'miner, zach' 'laffey, aaron' 'feldman, scott' 'nolasco, ricky'
 'marshall, sean' 'stults, eric' 'chavez, jesse' 'detwiler, ross'
 'kluber, corey' 'cecil, brett' 'parra, manny' 'hill, rich'
 'shields, james' 'garcía, jaime' 'harrell, lucas' 'volquez, edinson'
 'vargas, jason' 'weaver, jered' 'wilson, c.j.' 'medlen, kris'
 'fister, doug' 'matusz, brian' 'billingsley, chad' 'davis, wade'
 'gallardo, yovani' 'marcum, shaun' 'lester, jon' 'kendrick, kyle'
 'gorzelanny, tom' 'kennedy, ian' 'miller, andrew' 'wright, steven'
 'leblanc, wade' 'scherzer, max' 'huff, david' 'lincecum, tim'
 'buchholz, clay' 'morrow, brandon' 'richard, clayton' 'arrieta, jake'
 'villanueva, carlos' 'ohlendorf, ross' 'price, david' 'sánchez, jonatha
n'
 'noesí, hector' 'griffin, a.j.' 'cueto, johnny' 'bailey, homer'
 'harrison, matt' 'jurrjens, jair' "o'sullivan, sean" 'happ, j.a.'
 'hefner, jeremy' 'lynn, lance' 'volstad, chris' 'tomlin, josh'
 'lannan, john' 'slowey, kevin' 'humber, philip' 'hochevar, luke'
 'pelfrey, mike' 'romero, ricky' 'gonzález, gio' 'hughes, phil'
 'swarzak, anthony' 'hanson, tommy' 'estrada, marco' 'morales, franklin'
 'doubront, felix' 'nova, iván' 'ogando, alexi' 'chacín, jhoulys'
 'rogers, esmil' 'carrasco, carlos' 'anderson, brett' 'worley, vance'
 'ross, tyson' 'drabek, kyle' 'wood, travis' 'masterson, justin'
 'phelps, david' 'hellickson, jeremy' 'niese, jonathon' 'kershaw, clayto
n'
 'cashner, andrew' 'duensing, brian' 'hunter, tommy' 'miley, wade'
 'stammen, craig' 'garza, matt' 'kuroda, hiroki' 'matsuzaka, daisuke'
 'quintana, jose' 'tillman, chris' 'minor, mike' 'latos, mat'
 'norris, bud' 'archer, chris' 'gibson, kyle' 'locke, jeff'
 'mcallister, zach' 'cobb, alex' 'samardzija, jeff' 'leake, mike'
 'cahill, trevor' 'santiago, héctor' 'holland, derek' 'peacock, brad'
 'peralta, wily' 'nicasio, juan' 'darvish, yu' 'alvarez iii, henderson'
 'delgado, randall' 'salazar, danny' 'bettis, chad' 'bumgarner, madison'
 'collmenter, josh' 'duffy, danny' 'gee, dillon' 'harvey, matt'
 'moore, matt' 'nelson, jimmy' 'oberholtzer, brett' 'parker, jarrod'
 'pomeranz, drew' 'porcello, rick' 'sale, chris' 'zimmermann, jordan'
 'de la rosa, rubby' 'pérez, martín' 'teheran, julio' 'shoemaker, matt'
 'diamond, scott' 'chatwood, tyler' 'cole, gerrit' 'cosart, jarred'
```

```
'eovaldi, nathan' 'gray, sonny' 'koehler, tom' 'lyles, jordan'
'mchugh, collin' 'milone, tommy' 'odorizzi, jake' 'roark, tanner'
'strasburg, stephen' 'bauer, trevor' 'turner, jacob' 'beachy, brandon'
'iwakuma, hisashi' 'ryu, hyun jin' 'wheeler, zack' 'ventura, yordano'
'corbin, patrick' 'fiers, mike' 'mikolas, miles' 'miller, shelby'
'paxton, james' 'richards, garrett' 'skaggs, tyler' 'erlin, robbie'
'keuchel, dallas' 'nuño, vidal' 'straily, dan' 'gausman, kevin'
'smyly, drew' 'walker, taijuan' 'wacha, michael' 'chen, wei-yin'
'wood, alex']
```

In [129]:

```
"""
def remove_accents(input_str):
    nfkd_form = unicodedata.normalize('NFKD', input_str)
    return "".join([c for c in nfkd_form if not unicodedata.combining(

def clean_name(name):
    name = name.lower()
    name = remove_accents(name)
    name = re.sub(r'[-.]', '', name)
    name = re.sub(r'\s+', ' ', name).strip()
    return name

final_2013_df['player_name'] = final_2013_df['player_name'].apply(clea
"""
```

In [130]:

```
"""
# Convert 'player_name' from "last name, first name" to "first name la
final_2013_df['Name'] = final_2013_df['player_name'].apply(lambda x: '
"""
```

In [131]:

```
"""
# Now, you can perform the merge using 'Name' and 'season' as the keys
better_2013_df = pd.merge(final_2013_df,
                          cleaning_filtered_df[['Name', 'season', 'Age
                          on=['Name', 'season'],
                          how='left')
"""
```

In [133]:
```
1 #better_2013_df
```

Out[133]:

	pitcher	player_name	season	season_total_pitches	pitch_type	season_total_count_by
0	112526	colon, bartolo	2013	10929	CH	
1	112526	colon, bartolo	2013	10929	FF	
2	112526	colon, bartolo	2013	10929	SI	
3	112526	colon, bartolo	2013	10929	SL	
4	115629	hawkins, latroy	2013	3189	CH	
...	...	...	...	...	...	
1535	622072	wood, alex	2013	4803	FF	
1536	622072	wood, alex	2013	4803	IN	
1537	622072	wood, alex	2013	4803	KC	
1538	622072	wood, alex	2013	4803	PO	
1539	622072	wood, alex	2013	4803	SI	

1540 rows × 19 columns

In [134]:
```
#better_2013_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1540 entries, 0 to 1539
Data columns (total 19 columns):
 #   Column                            Non-Null Count  Dtype  
---  ------                            --------------  -----  
 0   pitcher                           1540 non-null   int64  
 1   player_name                       1540 non-null   object 
 2   season                            1540 non-null   int32  
 3   season_total_pitches              1540 non-null   int64  
 4   pitch_type                        1540 non-null   object 
 5   season_total_count_by_pitch_type  1540 non-null   int64  
 6   count_by_pitch_type               1540 non-null   int64  
 7   release_speed_weighted_avg        1540 non-null   float64
 8   release_pos_x_weighted_avg        1540 non-null   float64
 9   release_pos_z_weighted_avg        1540 non-null   float64
 10  vx0_weighted_avg                  1540 non-null   float64
 11  vy0_weighted_avg                  1540 non-null   float64
 12  vz0_weighted_avg                  1540 non-null   float64
 13  ax_weighted_avg                   1540 non-null   float64
 14  ay_weighted_avg                   1540 non-null   float64
 15  az_weighted_avg                   1540 non-null   float64
 16  release_pos_y_weighted_avg        1540 non-null   float64
 17  Name                              1540 non-null   object 
 18  Age                               1331 non-null   float64
dtypes: float64(11), int32(1), int64(4), object(3)
memory usage: 222.7+ KB
```

In [135]:
```
#better_2013_df.to_csv('data/better_2013_df.csv')
```

2014

In [136]:
```
#all_2014_stats_df = pd.read_csv('~/Documents/Flatiron/Project_5_/data/
```

In [137]:
```
"""
all_2014_stats_df.drop(columns=['batter', 'events', 'description', 'zo
                                'des', 'game_type', 'stand', 'home_tea
                                'away_team', 'type', 'hit_location', 'l
                                'balls', 'strikes', 'pfx_x', 'spin_dir
                                'pfx_z', 'plate_x', 'plate_z', 'on_3b'
                                'on_2b', 'on_1b', 'outs_when_up', 'inn
                                'inning_topbot', 'hc_x', 'hc_y', 'fiel
                                'umpire', 'sv_id', 'hit_distance_sc',
                                 'sz_bot', 'launch_speed', 'launch_ang
                                'pitcher.1', 'fielder_2.1', 'fielder_3
                                'fielder_5', 'fielder_6', 'fielder_7',
                                'fielder_9', 'estimated_ba_using_speed
                                'estimated_woba_using_speedangle', 'ba
                                'launch_speed_angle', 'woba_value', 'w
                                'at_bat_number', 'pitch_number', 'home_
                                'bat_score', 'fld_score', 'post_home_s
                                'post_fld_score', 'post_away_score', '
                                'of_fielding_alignment', 'delta_home_w
                                'delta_run_exp', 'spin_rate_deprecated
                               'break_length_deprecated', 'tfs_deprecat
all_2014_stats_df.head()
"""
```

Out[137]:

	pitch_type	game_date	release_speed	release_pos_x	release_pos_z	player_name	pitche
0	FF	2014-09-27	93.7	-1.62	6.28	Hawkins, LaTroy	11562
1	FF	2014-09-27	93.1	-1.63	6.18	Hawkins, LaTroy	11562
2	CU	2014-09-27	77.2	-1.53	6.58	Hawkins, LaTroy	11562
3	SL	2014-09-27	87.1	-1.61	6.38	Hawkins, LaTroy	11562
4	FF	2014-09-27	92.3	-1.57	6.10	Hawkins, LaTroy	11562

5 rows × 21 columns

In [138]:
```
"""
all_2014_stats_df.drop(columns=['spin_axis', 'effective_speed',
                                'release_spin_rate', 'release_extension
"""
```

In [139]:
```
#all_2014_stats_df = all_2014_stats_df.dropna(axis=0)
```

In [140]:
```
#all_2014_stats_df.reset_index(inplace=True)
```

In [141]:
```
1 #all_2014_stats_df.drop('index', axis=1)
```

Out[141]:

	pitch_type	game_date	release_speed	release_pos_x	release_pos_z	player_name
0	FF	2014-09-27	93.7	-1.62	6.28	Hawkins, LaTroy
1	FF	2014-09-27	93.1	-1.63	6.18	Hawkins, LaTroy
2	CU	2014-09-27	77.2	-1.53	6.58	Hawkins, LaTroy
3	SL	2014-09-27	87.1	-1.61	6.38	Hawkins, LaTroy
4	FF	2014-09-27	92.3	-1.57	6.10	Hawkins, LaTroy
...	...	...	...	...	...	...
452894	CH	2014-04-01	83.3	1.16	5.81	Wood, Alex
452895	SI	2014-04-01	91.5	0.89	5.88	Wood, Alex
452896	CH	2014-04-01	83.4	1.12	5.83	Wood, Alex
452897	SI	2014-04-01	92.0	1.03	5.83	Wood, Alex
452898	SI	2014-04-01	90.7	1.03	5.83	Wood, Alex

452899 rows × 17 columns

In [142]:

```
"""
# Group by 'game_date' and 'pitcher' to calculate the total pitches
total_pitches = all_2014_stats_df.groupby(['game_date', 'pitcher', 'pla

# Group by 'game_date', 'pitcher', and 'pitch_type' to calculate the su
total_pitches_by_type = all_2014_stats_df.groupby(['game_date', 'pitch

# Calculate averages of the specified metrics for each pitch type, grou
avg_metrics = all_2014_stats_df.groupby(['game_date', 'pitcher', 'play
    'release_speed': 'mean',
    'release_pos_x': 'mean',
    'release_pos_z': 'mean',
    'vx0': 'mean',
    'vy0': 'mean',
    'vz0': 'mean',
    'ax': 'mean',
    'ay': 'mean',
    'az': 'mean',
    'release_pos_y': 'mean',
}).reset_index()

grouped_2014_df = total_pitches.merge(total_pitches_by_type, on=['game_
grouped_2014_df = grouped_2014_df.merge(avg_metrics, on=['game_date',

grouped_2014_df.head()
"""
```

Out[142]:

	game_date	pitcher	player_name	total_pitches	pitch_type	count_by_pitch_type	rele
0	2014-03-22	450306	Vargas, Jason	96	CH	30	
1	2014-03-22	450306	Vargas, Jason	96	CU	14	
2	2014-03-22	450306	Vargas, Jason	96	FC	5	
3	2014-03-22	450306	Vargas, Jason	96	FF	46	
4	2014-03-22	450306	Vargas, Jason	96	SI	1	
...	...	...	...	...	...	...	
27564	2014-09-28	605135	Bassitt, Chris	97	FF	1	
27565	2014-09-28	605135	Bassitt, Chris	97	SI	43	
27566	2014-09-28	605135	Bassitt, Chris	97	SL	23	
27567	2014-09-28	608665	Graveman, Kendall	10	SI	9	
27568	2014-09-28	608665	Graveman, Kendall	10	SL	1	

27569 rows × 16 columns

In [143]:

```
"""
grouped_2014_df['game_date'] = pd.to_datetime(grouped_2014_df['game_da
grouped_2014_df['season'] = grouped_2014_df['game_date'].dt.year

# Step 1: Season Total Pitches
season_total_pitches = grouped_2014_df.groupby(['pitcher', 'player_nam

# Step 2: Season Total by Pitch Type
season_total_by_pitch_type = grouped_2014_df.groupby(['pitcher', 'play

# Weighted Averages Calculation Setup
weighted_avg_columns = ['release_speed', 'release_pos_x', 'release_pos_
for col in weighted_avg_columns:
    grouped_2014_df[f'{col}_product'] = grouped_2014_df[col] * grouped_

weighted_avg_aggregations = {f'{col}_product': 'sum' for col in weight
weighted_avg_aggregations['count_by_pitch_type'] = 'sum'

# Aggregate for weighted averages
weighted_avg_df = grouped_2014_df.groupby(['pitcher', 'player_name', '

# Calculate weighted averages
for col in weighted_avg_columns:
    weighted_avg_df[f'{col}_weighted_avg'] = weighted_avg_df[f'{col}_p

# Cleanup
weighted_avg_df.drop(columns=[f'{col}_product' for col in weighted_avg_

# Merge season totals and weighted averages
final_2014_df = pd.merge(season_total_pitches, season_total_by_pitch_ty
final_2014_df = pd.merge(final_2014_df, weighted_avg_df, on=['pitcher'

final_2014_df.head()
"""
```

Out[143]:

	pitcher	player_name	season	season_total_pitches	pitch_type	season_total_count_by_pit
0	112526	Colon, Bartolo	2014	12322	CH	
1	112526	Colon, Bartolo	2014	12322	FF	
2	112526	Colon, Bartolo	2014	12322	IN	
3	112526	Colon, Bartolo	2014	12322	SI	
4	112526	Colon, Bartolo	2014	12322	SL	

In [144]:

```
"""
final_2014_df['player_name'] = final_2014_df['player_name'].str.lower(
final_2014_df.head()
"""
```

Out[144]:

	pitcher	player_name	season	season_total_pitches	pitch_type	season_total_count_by
0	112526	colon, bartolo	2014	12322	CH	
1	112526	colon, bartolo	2014	12322	FF	
2	112526	colon, bartolo	2014	12322	IN	
3	112526	colon, bartolo	2014	12322	SI	
4	112526	colon, bartolo	2014	12322	SL	
...	...	...	...	...	...	
1447	628333	despaigne, odrisamer	2014	10794	FC	
1448	628333	despaigne, odrisamer	2014	10794	FF	
1449	628333	despaigne, odrisamer	2014	10794	PO	
1450	628333	despaigne, odrisamer	2014	10794	SI	
1451	628333	despaigne, odrisamer	2014	10794	SL	

1452 rows × 17 columns

In [145]:
```
1 #print(final_2014_df['player_name'].unique())
```

```
['colon, bartolo' 'hawkins, latroy' 'wright, jamey' 'chen, bruce'
 'wolf, randy' 'nathan, joe' 'burnett, a.j.' 'penny, brad' 'hudson, tim'
 'downs, scott' 'arroyo, bronson' 'benoit, joaquín' 'beckett, josh'
 'belisle, matt' 'buehrle, mark' 'sabathia, cc' 'vogelsong, ryan'
 'dickey, r.a.' 'affeldt, jeremy' 'lohse, kyle' 'lackey, john'
 'de la rosa, jorge' 'bedard, erik' 'lewis, colby' 'peavy, jake'
 'harang, aaron' 'pérez, oliver' 'guthrie, jeremy' 'williams, jerome'
 'capuano, chris' 'wainwright, adam' 'greinke, zack' 'floyd, gavin'
 'haren, dan' 'jackson, edwin' 'santana, ervin' 'correia, kevin'
 'simon, alfredo' 'mcgowan, dustin' 'germano, justin' 'maholm, paul'
 'cain, matt' 'hamels, cole' 'kazmir, scott' 'stauffer, tim'
 'young, chris' 'danks, john' 'hernandez, roberto' 'francis, jeff'
 'hernández, félix' 'petit, yusmeiro' 'verlander, justin'
 'liriano, francisco' 'saunders, joe' 'jiménez, ubaldo' 'hammel, jason'
 'rodriguez, wandy' 'sánchez, aníbal' 'rowland-smith, ryan' 'duke, zach'
 'mccarthy, brandon' 'feldman, scott' 'nolasco, ricky' 'marshall, sean'
 'stults, eric' 'chavez, jesse' 'eveland, dana' 'detwiler, ross'
 'kluber, corey' 'cecil, brett' 'parra, manny' 'hill, rich'
 'shields, james' 'garcía, jaime' 'harrell, lucas' 'volquez, edinson'
 'vargas, jason' 'weaver, jered' 'wilson, c.j.' 'fister, doug'
 'matusz, brian' 'davis, wade' 'gallardo, yovani' 'lester, jon'
 'kendrick, kyle' 'gorzelanny, tom' 'kennedy, ian' 'miller, andrew'
 'wright, steven' 'leblanc, wade' 'scherzer, max' 'huff, david'
 'lincecum, tim' 'buchholz, clay' 'morrow, brandon' 'arrieta, jake'
 'villanueva, carlos' 'price, david' 'sánchez, jonathan' 'noesí, hector'
 'cueto, johnny' 'bailey, homer' 'harrison, matt' 'jurrjens, jair'
"o'sullivan, sean" 'happ, j.a.' 'lynn, lance' 'tomlin, josh'
 'lannan, john' 'slowey, kevin' 'humber, philip' 'pelfrey, mike'
 'gonzález, gio' 'hughes, phil' 'swarzak, anthony' 'hanson, tommy'
 'estrada, marco' 'paulino, felipe' 'morales, franklin' 'doubront, felix'
 'nova, iván' 'ogando, alexi' 'chacín, jhoulys' 'rogers, esmil'
 'carrasco, carlos' 'anderson, brett' 'worley, vance' 'ross, tyson'
 'drabek, kyle' 'wood, travis' 'masterson, justin' 'phelps, david'
 'hellickson, jeremy' 'niese, jonathon' 'kershaw, clayton'
 'cashner, andrew' 'duensing, brian' 'hunter, tommy' 'miley, wade'
 'stammen, craig' 'garza, matt' 'kuroda, hiroki' 'matsuzaka, daisuke'
 'quintana, jose' 'pineda, michael' 'tillman, chris' 'minor, mike'
 'latos, mat' 'norris, bud' 'archer, chris' 'gibson, kyle' 'locke, jeff'
 'mcallister, zach' 'cobb, alex' 'samardzija, jeff' 'leake, mike'
 'cahill, trevor' 'santiago, héctor' 'anderson, chase' 'holland, derek'
 'peacock, brad' 'peralta, wily' 'nicasio, juan' 'darvish, yu'
 'alvarez iii, henderson' 'delgado, randall' 'salazar, danny'
 'bettis, chad' 'bumgarner, madison' 'collmenter, josh' 'duffy, danny'
 'gee, dillon' 'moore, matt' 'nelson, jimmy' 'oberholtzer, brett'
 'pomeranz, drew' 'porcello, rick' 'sale, chris' 'zimmermann, jordan'
 'de la rosa, rubby' 'pérez, martín' 'teheran, julio' 'shoemaker, matt'
 'chatwood, tyler' 'cole, gerrit' 'cosart, jarred' 'desclafani, anthony'
 'eovaldi, nathan' 'gray, sonny' 'hendricks, kyle' 'hudson, daniel'
 'koehler, tom' 'lyles, jordan' 'mchugh, collin' 'milone, tommy'
 'odorizzi, jake' 'roark, tanner' 'strasburg, stephen' 'bauer, trevor'
 'turner, jacob' 'iwakuma, hisashi' 'tanaka, masahiro' 'ryu, hyun jin'
 'wheeler, zack' 'ventura, yordano' 'alvarez, r.j.' 'fiers, mike'
 'heaney, andrew' 'hutchison, drew' 'mikolas, miles' 'miller, shelby'
 'paxton, james' 'richards, garrett' 'skaggs, tyler' 'erlin, robbie'
 'keuchel, dallas' 'nuño, vidal' 'straily, dan' 'stroman, marcus'
 'foltynewicz, mike' 'gausman, kevin' 'ray, robbie' 'sanchez, aaron'
 'smyly, drew' 'walker, taijuan' 'degrom, jacob' 'gonzales, marco'
```

```
'norris, daniel' 'bassitt, chris' 'wisler, matt' 'elías, roenis'
'wacha, michael' 'graveman, kendall' 'chen, wei-yin' 'wood, alex'
'despaigne, odrisamer']
```

In [146]:

```
"""
def remove_accents(input_str):
    nfkd_form = unicodedata.normalize('NFKD', input_str)
    return "".join([c for c in nfkd_form if not unicodedata.combining(

def clean_name(name):
    name = name.lower()
    name = remove_accents(name)
    name = re.sub(r'[-.]', '', name)
    name = re.sub(r'\s+', ' ', name).strip()
    return name

final_2014_df['player_name'] = final_2014_df['player_name'].apply(clea
"""
```

In [147]:

```
"""
# Convert 'player_name' from "last name, first name" to "first name la
final_2014_df['Name'] = final_2014_df['player_name'].apply(lambda x: '
"""
```

In [148]:

```
"""
# Now, you can perform the merge using 'Name' and 'season' as the keys
better_2014_df = pd.merge(final_2014_df,
                          cleaning_filtered_df[['Name', 'season', 'Age
                          on=['Name', 'season'],
                          how='left')
"""
```

In [151]:

```
#better_2014_df.head()
```

Out[151]:

	pitcher	player_name	season	season_total_pitches	pitch_type	season_total_count_by_pit
0	112526	colon, bartolo	2014	12322	CH	
1	112526	colon, bartolo	2014	12322	FF	
2	112526	colon, bartolo	2014	12322	IN	
3	112526	colon, bartolo	2014	12322	SI	
4	112526	colon, bartolo	2014	12322	SL	

In [152]:
```
#better_2014_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1497 entries, 0 to 1496
Data columns (total 19 columns):
 #   Column                            Non-Null Count  Dtype
---  ------                            --------------  -----
 0   pitcher                           1497 non-null   int64
 1   player_name                       1497 non-null   object
 2   season                            1497 non-null   int32
 3   season_total_pitches              1497 non-null   int64
 4   pitch_type                        1497 non-null   object
 5   season_total_count_by_pitch_type  1497 non-null   int64
 6   count_by_pitch_type               1497 non-null   int64
 7   release_speed_weighted_avg        1497 non-null   float64
 8   release_pos_x_weighted_avg        1497 non-null   float64
 9   release_pos_z_weighted_avg        1497 non-null   float64
 10  vx0_weighted_avg                  1497 non-null   float64
 11  vy0_weighted_avg                  1497 non-null   float64
 12  vz0_weighted_avg                  1497 non-null   float64
 13  ax_weighted_avg                   1497 non-null   float64
 14  ay_weighted_avg                   1497 non-null   float64
 15  az_weighted_avg                   1497 non-null   float64
 16  release_pos_y_weighted_avg        1497 non-null   float64
 17  Name                              1497 non-null   object
 18  Age                               1297 non-null   float64
dtypes: float64(11), int32(1), int64(4), object(3)
memory usage: 216.5+ KB
```

In [153]:
```
#better_2014_df.to_csv('data/better_2014_df.csv')
```

2015

In [154]:
```
#all_2015_stats_df = pd.read_csv('~/Documents/Flatiron/Project_5_/data/
```

In [155]:

```
"""
all_2015_stats_df.drop(columns=['batter', 'events', 'description', 'zo
                                'des', 'game_type', 'stand', 'home_tea
                                'away_team', 'type', 'hit_location', 'l
                                'balls', 'strikes', 'pfx_x', 'spin_dir
                                'pfx_z', 'plate_x', 'plate_z', 'on_3b'
                                'on_2b', 'on_1b', 'outs_when_up', 'inn
                                'inning_topbot', 'hc_x', 'hc_y', 'fiel
                                'umpire', 'sv_id', 'hit_distance_sc',
                                 'sz_bot', 'launch_speed', 'launch_ang
                                'pitcher.1', 'fielder_2.1', 'fielder_3
                                'fielder_5', 'fielder_6', 'fielder_7',
                                'fielder_9', 'estimated_ba_using_speed
                                'estimated_woba_using_speedangle', 'bal
                                'launch_speed_angle', 'woba_value', 'w
                                'at_bat_number', 'pitch_number', 'home_
                                'bat_score', 'fld_score', 'post_home_s
                                'post_fld_score', 'post_away_score', '
                                'of_fielding_alignment', 'delta_home_w
                                'delta_run_exp', 'spin_rate_deprecated
                               'break_length_deprecated', 'tfs_deprecat
all_2015_stats_df.head()
"""
```

Out[155]:

	pitch_type	game_date	release_speed	release_pos_x	release_pos_z	player_name	pitche
0	FF	2015-10-03	92.8	-1.30	6.52	Hawkins, LaTroy	11562
1	FF	2015-10-03	92.2	-1.05	6.71	Hawkins, LaTroy	11562
2	SL	2015-10-03	86.8	-1.79	6.49	Hawkins, LaTroy	11562
3	SL	2015-09-30	86.0	-1.75	6.31	Hawkins, LaTroy	11562
4	FF	2015-09-30	94.1	-1.74	6.29	Hawkins, LaTroy	11562

5 rows × 21 columns

In [156]:

```
"""
all_2015_stats_df.drop(columns=['spin_axis', 'effective_speed',
                                'release_spin_rate', 'release_extension
"""
```

In [157]:

```
#all_2015_stats_df = all_2015_stats_df.dropna(axis=0)
```

In [158]:

```
#all_2015_stats_df.reset_index(inplace=True)
```

In [159]:
```
#all_2015_stats_df.drop('index', axis=1)
```

Out[159]:

	pitch_type	game_date	release_speed	release_pos_x	release_pos_z	player_name
0	FF	2015-10-03	92.8	-1.30	6.52	Hawkins, LaTroy
1	FF	2015-10-03	92.2	-1.05	6.71	Hawkins, LaTroy
2	SL	2015-10-03	86.8	-1.79	6.49	Hawkins, LaTroy
3	SL	2015-09-30	86.0	-1.75	6.31	Hawkins, LaTroy
4	FF	2015-09-30	94.1	-1.74	6.29	Hawkins, LaTroy
...	...	...	...	...	...	...
421311	SI	2015-04-07	90.3	1.94	5.71	Wood, Alex
421312	SI	2015-04-07	90.4	2.42	5.47	Wood, Alex
421313	SI	2015-04-07	91.3	2.33	5.45	Wood, Alex
421314	SI	2015-04-07	89.7	2.22	5.55	Wood, Alex
421315	SI	2015-04-07	90.4	2.42	5.63	Wood, Alex

421316 rows × 17 columns

In [160]:

```
"""
# Group by 'game_date' and 'pitcher' to calculate the total pitches
total_pitches = all_2015_stats_df.groupby(['game_date', 'pitcher', 'pla

# Group by 'game_date', 'pitcher', and 'pitch_type' to calculate the su
total_pitches_by_type = all_2015_stats_df.groupby(['game_date', 'pitch

# Calculate averages of the specified metrics for each pitch type, grou
avg_metrics = all_2015_stats_df.groupby(['game_date', 'pitcher', 'play
    'release_speed': 'mean',
    'release_pos_x': 'mean',
    'release_pos_z': 'mean',
    'vx0': 'mean',
    'vy0': 'mean',
    'vz0': 'mean',
    'ax': 'mean',
    'ay': 'mean',
    'az': 'mean',
    'release_pos_y': 'mean',
}).reset_index()

grouped_2015_df = total_pitches.merge(total_pitches_by_type, on=['game_
grouped_2015_df = grouped_2015_df.merge(avg_metrics, on=['game_date',

grouped_2015_df.head()
"""
```

Out[160]:

	game_date	pitcher	player_name	total_pitches	pitch_type	count_by_pitch_type	release_
0	2015-04-05	425794	Wainwright, Adam	101	CH	2	83.7
1	2015-04-05	425794	Wainwright, Adam	101	CU	26	75.9
2	2015-04-05	425794	Wainwright, Adam	101	FC	46	86.5
3	2015-04-05	425794	Wainwright, Adam	101	FF	23	89.9
4	2015-04-05	425794	Wainwright, Adam	101	SI	4	89.3

In [161]:

```
"""
grouped_2015_df['game_date'] = pd.to_datetime(grouped_2015_df['game_da
grouped_2015_df['season'] = grouped_2015_df['game_date'].dt.year

# Step 1: Season Total Pitches
season_total_pitches = grouped_2015_df.groupby(['pitcher', 'player_nam

# Step 2: Season Total by Pitch Type
season_total_by_pitch_type = grouped_2015_df.groupby(['pitcher', 'play

# Weighted Averages Calculation Setup
weighted_avg_columns = ['release_speed', 'release_pos_x', 'release_pos_
for col in weighted_avg_columns:
    grouped_2015_df[f'{col}_product'] = grouped_2015_df[col] * grouped_

weighted_avg_aggregations = {f'{col}_product': 'sum' for col in weight
weighted_avg_aggregations['count_by_pitch_type'] = 'sum'

# Aggregate for weighted averages
weighted_avg_df = grouped_2015_df.groupby(['pitcher', 'player_name', '

# Calculate weighted averages
for col in weighted_avg_columns:
    weighted_avg_df[f'{col}_weighted_avg'] = weighted_avg_df[f'{col}_p

# Cleanup
weighted_avg_df.drop(columns=[f'{col}_product' for col in weighted_avg_

# Merge season totals and weighted averages
final_2015_df = pd.merge(season_total_pitches, season_total_by_pitch_ty
final_2015_df = pd.merge(final_2015_df, weighted_avg_df, on=['pitcher'

final_2015_df.head()
"""
```

Out[161]:

	pitcher	player_name	season	season_total_pitches	pitch_type	season_total_count_by_pit
0	112526	Colon, Bartolo	2015	11167	CH	
1	112526	Colon, Bartolo	2015	11167	CU	
2	112526	Colon, Bartolo	2015	11167	FF	
3	112526	Colon, Bartolo	2015	11167	IN	
4	112526	Colon, Bartolo	2015	11167	SI	

In [162]:

```
"""
final_2015_df['player_name'] = final_2015_df['player_name'].str.lower(
final_2015_df.head()
"""
```

Out[162]:

	pitcher	player_name	season	season_total_pitches	pitch_type	season_total_count_by_pit
0	112526	colon, bartolo	2015	11167	CH	
1	112526	colon, bartolo	2015	11167	CU	
2	112526	colon, bartolo	2015	11167	FF	
3	112526	colon, bartolo	2015	11167	IN	
4	112526	colon, bartolo	2015	11167	SI	

In [163]:
```
1 #print(final_2015_df['player_name'].unique())
```

```
['colon, bartolo' 'hawkins, latroy' 'chen, bruce' 'wolf, randy'
 'nathan, joe' 'marquis, jason' 'burnett, a.j.' 'zito, barry'
 'hudson, tim' 'benoit, joaquín' 'belisle, matt' 'buehrle, mark'
 'sabathia, cc' 'vogelsong, ryan' 'dickey, r.a.' 'affeldt, jeremy'
 'lohse, kyle' 'lackey, john' 'de la rosa, jorge' 'lewis, colby'
 'peavy, jake' 'harang, aaron' 'pérez, oliver' 'guthrie, jeremy'
 'williams, jerome' 'capuano, chris' 'wainwright, adam' 'greinke, zack'
 'floyd, gavin' 'haren, dan' 'jackson, edwin' 'santana, ervin'
 'narveson, chris' 'correia, kevin' 'simon, alfredo' 'blanton, joe'
 'mcgowan, dustin' 'cain, matt' 'hamels, cole' 'kazmir, scott'
 'stauffer, tim' 'young, chris' 'danks, john' 'hernandez, roberto'
 'francis, jeff' 'hernández, félix' 'petit, yusmeiro' 'verlander, justin'
 'liriano, francisco' 'jiménez, ubaldo' 'hammel, jason' 'rodriguez, wand
y'
 'sánchez, aníbal' 'davies, kyle' 'duke, zach' 'mccarthy, brandon'
 'laffey, aaron' 'feldman, scott' 'nolasco, ricky' 'stults, eric'
 'chavez, jesse' 'eveland, dana' 'detwiler, ross' 'kluber, corey'
 'cecil, brett' 'parra, manny' 'hill, rich' 'shields, james'
 'garcía, jaime' 'guerra, junior' 'volquez, edinson' 'vargas, jason'
 'weaver, jered' 'wilson, c.j.' 'medlen, kris' 'fister, doug'
 'matusz, brian' 'billingsley, chad' 'davis, wade' 'gallardo, yovani'
 'marcum, shaun' 'lester, jon' 'kendrick, kyle' 'gorzelanny, tom'
 'kennedy, ian' 'miller, andrew' 'wright, steven' 'scherzer, max'
 'huff, david' 'lincecum, tim' 'buchholz, clay' 'morrow, brandon'
 'richard, clayton' 'arrieta, jake' 'villanueva, carlos' 'ohlendorf, ros
s'
 'price, david' 'noesí, hector' 'cueto, johnny' 'bailey, homer'
 'harrison, matt' 'reyes, jo-jo' "o'sullivan, sean" 'happ, j.a.'
 'lynn, lance' 'volstad, chris' 'tomlin, josh' 'hochevar, luke'
 'pelfrey, mike' 'gonzález, gio' 'hughes, phil' 'swarzak, anthony'
 'estrada, marco' 'morales, franklin' 'doubront, felix' 'nova, iván'
 'ogando, alexi' 'chacín, jhoulys' 'rogers, esmil' 'carrasco, carlos'
 'anderson, brett' 'worley, vance' 'ross, tyson' 'drabek, kyle'
 'wood, travis' 'masterson, justin' 'phelps, david' 'hellickson, jeremy'
 'niese, jonathon' 'kershaw, clayton' 'cashner, andrew' 'duensing, brian'
 'hunter, tommy' 'miley, wade' 'stammen, craig' 'garza, matt'
 'quintana, jose' 'pineda, michael' 'tillman, chris' 'latos, mat'
 'norris, bud' 'archer, chris' 'gibson, kyle' 'locke, jeff'
 'mcallister, zach' 'samardzija, jeff' 'leake, mike' 'cahill, trevor'
 'santiago, héctor' 'anderson, chase' 'holland, derek' 'peacock, brad'
 'peralta, wily' 'nicasio, juan' 'alvarez iii, henderson'
 'delgado, randall' 'salazar, danny' 'bettis, chad' 'bumgarner, madison'
 'collmenter, josh' 'duffy, danny' 'gee, dillon' 'harvey, matt'
 'moore, matt' 'nelson, jimmy' 'oberholtzer, brett' 'pomeranz, drew'
 'porcello, rick' 'sale, chris' 'zimmermann, jordan' 'de la rosa, rubby'
 'pérez, martín' 'teheran, julio' 'shoemaker, matt' 'andriese, matt'
 'cole, gerrit' 'cosart, jarred' 'desclafani, anthony' 'eovaldi, nathan'
 'gray, sonny' 'hendricks, kyle' 'hudson, daniel' 'koehler, tom'
 'lyles, jordan' 'mchugh, collin' 'milone, tommy' 'montgomery, mike'
 'odorizzi, jake' 'roark, tanner' 'strasburg, stephen' 'bauer, trevor'
 'beachy, brandon' 'lorenzen, michael' 'iwakuma, hisashi'
 'tanaka, masahiro' 'ureña, josé' 'ventura, yordano' 'alvarez, r.j.'
 'boyd, matthew' 'corbin, patrick' 'fiers, mike' 'heaney, andrew'
 'hutchison, drew' 'matz, steven' 'miller, shelby' 'paxton, james'
 'richards, garrett' 'erlin, robbie' 'keuchel, dallas' 'nuño, vidal'
 'straily, dan' 'stroman, marcus' 'foltynewicz, mike' 'gausman, kevin'
 'gray, jon' 'ray, robbie' 'sanchez, aaron' 'smyly, drew'
```

```
'syndergaard, noah' 'velasquez, vince' 'walker, taijuan'
'montas, frankie' 'rodriguez, eduardo' 'degrom, jacob' 'gonzales, marco'
'eickhoff, jerad' 'norris, daniel' 'bassitt, chris' 'davies, zach'
'houser, adrian' 'lópez, jorge' 'nola, aaron' 'ross, joe' 'wisler, matt'
'elías, roenis' 'rodón, carlos' 'wacha, michael' 'graveman, kendall'
'chen, wei-yin' 'mccullers jr., lance' 'wood, alex' 'severino, luis'
'despaigne, odrisamer' 'godley, zack']
```

In [164]:

```
"""
def remove_accents(input_str):
    nfkd_form = unicodedata.normalize('NFKD', input_str)
    return "".join([c for c in nfkd_form if not unicodedata.combining(

def clean_name(name):
    name = name.lower()
    name = remove_accents(name)
    name = re.sub(r'[-.]', '', name)
    name = re.sub(r'\s+', ' ', name).strip()
    return name

final_2015_df['player_name'] = final_2015_df['player_name'].apply(clea
"""
```

In [165]:

```
"""
# Convert 'player_name' from "last name, first name" to "first name la
final_2015_df['Name'] = final_2015_df['player_name'].apply(lambda x: '
"""
```

In [166]:

```
"""
# Now, you can perform the merge using 'Name' and 'season' as the keys
better_2015_df = pd.merge(final_2015_df,
                          cleaning_filtered_df[['Name', 'season', 'Age
                          on=['Name', 'season'],
                          how='left')
"""
```

In [167]:

```
#better_2015_df.head()
```

Out[167]:

	pitcher	player_name	season	season_total_pitches	pitch_type	season_total_count_by_pit
0	112526	colon, bartolo	2015	11167	CH	
1	112526	colon, bartolo	2015	11167	CU	
2	112526	colon, bartolo	2015	11167	FF	
3	112526	colon, bartolo	2015	11167	IN	
4	112526	colon, bartolo	2015	11167	SI	

In [168]:
```
1 #better_2015_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1475 entries, 0 to 1474
Data columns (total 19 columns):
 #   Column                            Non-Null Count  Dtype
---  ------                            --------------  -----
 0   pitcher                           1475 non-null   int64
 1   player_name                       1475 non-null   object
 2   season                            1475 non-null   int32
 3   season_total_pitches              1475 non-null   int64
 4   pitch_type                        1475 non-null   object
 5   season_total_count_by_pitch_type  1475 non-null   int64
 6   count_by_pitch_type               1475 non-null   int64
 7   release_speed_weighted_avg        1475 non-null   float64
 8   release_pos_x_weighted_avg        1475 non-null   float64
 9   release_pos_z_weighted_avg        1475 non-null   float64
 10  vx0_weighted_avg                  1475 non-null   float64
 11  vy0_weighted_avg                  1475 non-null   float64
 12  vz0_weighted_avg                  1475 non-null   float64
 13  ax_weighted_avg                   1475 non-null   float64
 14  ay_weighted_avg                   1475 non-null   float64
 15  az_weighted_avg                   1475 non-null   float64
 16  release_pos_y_weighted_avg        1475 non-null   float64
 17  Name                              1475 non-null   object
 18  Age                               1276 non-null   float64
dtypes: float64(11), int32(1), int64(4), object(3)
memory usage: 213.3+ KB
```

In [169]:
```
1 #better_2015_df.to_csv('data/better_2015_df.csv')
```

2016

In [279]:
```
1 #all_2016_stats_df = pd.read_csv('~/Documents/Flatiron/Project_5_/data/
```

In [280]:
```
"""
all_2016_stats_df.drop(columns=['batter', 'events', 'description', 'zo
                                'des', 'game_type', 'stand', 'home_tea
                                'away_team', 'type', 'hit_location', '
                                'balls', 'strikes', 'pfx_x', 'spin_dir
                                'pfx_z', 'plate_x', 'plate_z', 'on_3b'
                                'on_2b', 'on_1b', 'outs_when_up', 'inn
                                'inning_topbot', 'hc_x', 'hc_y', 'fiel
                                'umpire', 'sv_id', 'hit_distance_sc',
                                 'sz_bot', 'launch_speed', 'launch_ang
                                'pitcher.1', 'fielder_2.1', 'fielder_3
                                'fielder_5', 'fielder_6', 'fielder_7',
                                'fielder_9', 'estimated_ba_using_speed
                                'estimated_woba_using_speedangle', 'ba
                                'launch_speed_angle', 'woba_value', 'w
                                'at_bat_number', 'pitch_number', 'home_
                                'bat_score', 'fld_score', 'post_home_s
                                'post_fld_score', 'post_away_score', '
                                'of_fielding_alignment', 'delta_home_w
                                'delta_run_exp', 'spin_rate_deprecated
                               'break_length_deprecated', 'tfs_deprec
all_2016_stats_df.head()
"""
```

Out[280]:

	pitch_type	game_date	release_speed	release_pos_x	release_pos_z	player_name	pitche
0	SI	2016-09-27	91.5	-2.44	6.38	Nathan, Joe	15027
1	CU	2016-09-27	80.6	-2.11	6.73	Nathan, Joe	15027
2	FF	2016-09-27	91.5	-1.98	6.80	Nathan, Joe	15027
3	FF	2016-09-27	91.4	-2.10	6.78	Nathan, Joe	15027
4	CU	2016-09-27	79.3	-2.02	6.99	Nathan, Joe	15027

5 rows × 21 columns

In [281]:
```
"""
all_2016_stats_df.drop(columns=['spin_axis', 'effective_speed',
                                'release_spin_rate', 'release_extension
"""
```

In [282]:
```
#all_2016_stats_df = all_2016_stats_df.dropna(axis=0)
```

In [283]:
```
#all_2016_stats_df.reset_index(inplace=True)
```

In [284]:
```
1 #all_2016_stats_df.drop('index', axis=1)
```

Out[284]:

	pitch_type	game_date	release_speed	release_pos_x	release_pos_z	player_name
0	SI	2016-09-27	91.5	-2.44	6.38	Nathan, Joe
1	CU	2016-09-27	80.6	-2.11	6.73	Nathan, Joe
2	FF	2016-09-27	91.5	-1.98	6.80	Nathan, Joe
3	FF	2016-09-27	91.4	-2.10	6.78	Nathan, Joe
4	CU	2016-09-27	79.3	-2.02	6.99	Nathan, Joe
...	...	...	...	...	...	...
428307	SI	2016-04-07	92.5	1.71	5.76	Wood, Alex
428308	KC	2016-04-07	84.7	1.81	5.62	Wood, Alex
428309	SI	2016-04-07	91.5	1.69	5.82	Wood, Alex
428310	SI	2016-04-07	92.5	1.70	5.69	Wood, Alex
428311	SI	2016-04-07	92.0	1.81	5.82	Wood, Alex

428312 rows × 17 columns

In [285]:

```
"""
# Group by 'game_date' and 'pitcher' to calculate the total pitches
total_pitches = all_2016_stats_df.groupby(['game_date', 'pitcher', 'pla

# Group by 'game_date', 'pitcher', and 'pitch_type' to calculate the su
total_pitches_by_type = all_2016_stats_df.groupby(['game_date', 'pitch

# Calculate averages of the specified metrics for each pitch type, grou
avg_metrics = all_2016_stats_df.groupby(['game_date', 'pitcher', 'play
    'release_speed': 'mean',
    'release_pos_x': 'mean',
    'release_pos_z': 'mean',
    'vx0': 'mean',
    'vy0': 'mean',
    'vz0': 'mean',
    'ax': 'mean',
    'ay': 'mean',
    'az': 'mean',
    'release_pos_y': 'mean',
}).reset_index()

grouped_2016_df = total_pitches.merge(total_pitches_by_type, on=['game_
grouped_2016_df = grouped_2016_df.merge(avg_metrics, on=['game_date',

grouped_2016_df.head()
"""
```

Out[285]:

	game_date	pitcher	player_name	total_pitches	pitch_type	count_by_pitch_type	release_
0	2016-04-03	112526	Colon, Bartolo	20	CH	2	81.8
1	2016-04-03	112526	Colon, Bartolo	20	FF	6	90.5
2	2016-04-03	112526	Colon, Bartolo	20	SI	12	88.6
3	2016-04-03	425794	Wainwright, Adam	96	CH	3	84.1
4	2016-04-03	425794	Wainwright, Adam	96	CU	36	75.3

In [286]:

```
"""
grouped_2016_df['game_date'] = pd.to_datetime(grouped_2016_df['game_da
grouped_2016_df['season'] = grouped_2016_df['game_date'].dt.year

# Step 1: Season Total Pitches
season_total_pitches = grouped_2016_df.groupby(['pitcher', 'player_nam

# Step 2: Season Total by Pitch Type
season_total_by_pitch_type = grouped_2016_df.groupby(['pitcher', 'play

# Weighted Averages Calculation Setup
weighted_avg_columns = ['release_speed', 'release_pos_x', 'release_pos_
for col in weighted_avg_columns:
    grouped_2016_df[f'{col}_product'] = grouped_2016_df[col] * grouped_

weighted_avg_aggregations = {f'{col}_product': 'sum' for col in weight
weighted_avg_aggregations['count_by_pitch_type'] = 'sum'

# Aggregate for weighted averages
weighted_avg_df = grouped_2016_df.groupby(['pitcher', 'player_name', '

# Calculate weighted averages
for col in weighted_avg_columns:
    weighted_avg_df[f'{col}_weighted_avg'] = weighted_avg_df[f'{col}_p

# Cleanup
weighted_avg_df.drop(columns=[f'{col}_product' for col in weighted_avg_

# Merge season totals and weighted averages
final_2016_df = pd.merge(season_total_pitches, season_total_by_pitch_ty
final_2016_df = pd.merge(final_2016_df, weighted_avg_df, on=['pitcher'

final_2016_df.head()
"""
```

Out[286]:

	pitcher	player_name	season	season_total_pitches	pitch_type	season_total_count_by_pit
0	112526	Colon, Bartolo	2016	11394	CH	
1	112526	Colon, Bartolo	2016	11394	FF	
2	112526	Colon, Bartolo	2016	11394	IN	
3	112526	Colon, Bartolo	2016	11394	SI	
4	112526	Colon, Bartolo	2016	11394	SL	

In [287]:

```
"""
final_2016_df['player_name'] = final_2016_df['player_name'].str.lower(
final_2016_df.head()
"""
```

Out[287]:

	pitcher	player_name	season	season_total_pitches	pitch_type	season_total_count_by_pit
0	112526	colon, bartolo	2016	11394	CH	
1	112526	colon, bartolo	2016	11394	FF	
2	112526	colon, bartolo	2016	11394	IN	
3	112526	colon, bartolo	2016	11394	SI	
4	112526	colon, bartolo	2016	11394	SL	

In [288]:

```
1 #print(final_2016_df['player_name'].unique())
```

```
['colon, bartolo' 'nathan, joe' 'benoit, joaquín' 'belisle, matt'
 'sabathia, cc' 'vogelsong, ryan' 'dickey, r.a.' 'lohse, kyle'
 'lackey, john' 'de la rosa, jorge' 'lewis, colby' 'peavy, jake'
 'pérez, oliver' 'wang, chien-ming' 'williams, jerome' 'capuano, chris'
 'wainwright, adam' 'greinke, zack' 'floyd, gavin' 'jackson, edwin'
 'santana, ervin' 'narveson, chris' 'simon, alfredo' 'blanton, joe'
 'mcgowan, dustin' 'cain, matt' 'hamels, cole' 'kazmir, scott'
 'young, chris' 'danks, john' 'hernandez, roberto' 'hernández, félix'
 'petit, yusmeiro' 'verlander, justin' 'liriano, francisco'
 'jiménez, ubaldo' 'hammel, jason' 'sánchez, aníbal' 'duke, zach'
 'mccarthy, brandon' 'feldman, scott' 'nolasco, ricky' 'chavez, jesse'
 'eveland, dana' 'detwiler, ross' 'kluber, corey' 'cecil, brett'
 'hill, rich' 'shields, james' 'garcía, jaime' 'guerra, junior'
 'harrell, lucas' 'volquez, edinson' 'vargas, jason' 'weaver, jered'
 'medlen, kris' 'fister, doug' 'matusz, brian' 'davis, wade'
 'gallardo, yovani' 'lester, jon' 'gorzelanny, tom' 'kennedy, ian'
 'miller, andrew' 'wright, steven' 'leblanc, wade' 'scherzer, max'
 'huff, david' 'lincecum, tim' 'buchholz, clay' 'morrow, brandon'
 'richard, clayton' 'arrieta, jake' 'villanueva, carlos' 'ohlendorf, ros
s'
 'price, david' 'griffin, a.j.' 'cueto, johnny' 'bailey, homer'
 'reyes, jo-jo' "o'sullivan, sean" 'happ, j.a.' 'luebke, cory'
 'tomlin, josh' 'hochevar, luke' 'pelfrey, mike' 'gonzález, gio'
 'hughes, phil' 'swarzak, anthony' 'estrada, marco' 'morales, franklin'
 'nova, iván' 'ogando, alexi' 'chacín, jhoulys' 'carrasco, carlos'
 'anderson, brett' 'worley, vance' 'ross, tyson' 'drabek, kyle'
 'wood, travis' 'phelps, david' 'hellickson, jeremy' 'niese, jonathon'
 'kershaw, clayton' 'cashner, andrew' 'duensing, brian' 'hunter, tommy'
 'miley, wade' 'garza, matt' 'quintana, jose' 'pineda, michael'
 'tillman, chris' 'latos, mat' 'norris, bud' 'archer, chris'
 'gibson, kyle' 'locke, jeff' 'mcallister, zach' 'cobb, alex'
 'samardzija, jeff' 'leake, mike' 'cahill, trevor' 'santiago, héctor'
 'anderson, chase' 'holland, derek' 'peacock, brad' 'peralta, wily'
 'nicasio, juan' 'darvish, yu' 'delgado, randall' 'salazar, danny'
 'bettis, chad' 'bumgarner, madison' 'collmenter, josh' 'duffy, danny'
 'gee, dillon' 'harvey, matt' 'moore, matt' 'nelson, jimmy'
 'oberholtzer, brett' 'pomeranz, drew' 'porcello, rick' 'sale, chris'
 'zimmermann, jordan' 'de la rosa, rubby' 'pérez, martín' 'teheran, juli
o'
 'shoemaker, matt' 'diamond, scott' 'anderson, tyler' 'andriese, matt'
 'chatwood, tyler' 'cole, gerrit' 'cosart, jarred' 'desclafani, anthony'
 'eovaldi, nathan' 'gray, sonny' 'hendricks, kyle' 'hudson, daniel'
 'koehler, tom' 'lyles, jordan' 'mchugh, collin' 'milone, tommy'
 'montgomery, mike' 'odorizzi, jake' 'roark, tanner' 'strasburg, stephen'
 'bauer, trevor' 'turner, jacob' 'lorenzen, michael' 'iwakuma, hisashi'
 'tanaka, masahiro' 'ryu, hyun jin' 'stripling, ross' 'ureña, josé'
 'ventura, yordano' 'boyd, matthew' 'corbin, patrick' 'fiers, mike'
 'heaney, andrew' 'hutchison, drew' 'matz, steven' 'miller, shelby'
 'paxton, james' 'richards, garrett' 'skaggs, tyler' 'erlin, robbie'
 'keuchel, dallas' 'nuño, vidal' 'straily, dan' 'stroman, marcus'
 'foltynewicz, mike' 'gausman, kevin' 'gray, jon' 'ray, robbie'
 'sanchez, aaron' 'smyly, drew' 'syndergaard, noah' 'taillon, jameson'
 'velasquez, vince' 'walker, taijuan' 'williams, trevor'
 'rodriguez, eduardo' 'degrom, jacob' 'eickhoff, jerad' 'norris, daniel'
 'weaver, luke' 'bassitt, chris' 'bundy, dylan' 'clevinger, mike'
 'davies, zach' 'fulmer, michael' 'musgrove, joe' 'nola, aaron'
 'ross, joe' 'snell, blake' 'wisler, matt' 'perdomo, luis' 'elías, roeni
```

```
s'
 'rodón, carlos' 'glasnow, tyler' 'lugo, seth' 'giolito, lucas'
 'wacha, michael' 'márquez, germán' 'graveman, kendall' 'stratton, chris'
 'suter, brent' 'chen, wei-yin' 'eflin, zach' 'mccullers jr., lance'
 'berríos, josé' 'blach, ty' 'wood, alex' 'severino, luis'
 'lópez, reynaldo' 'maeda, kenta' 'despaigne, odrisamer' 'urías, julio'
 'manaea, sean' 'kuhl, chad' 'brault, steven' 'godley, zack'
 'hoffman, jeff']
```

In [289]:

```
"""
def remove_accents(input_str):
    nfkd_form = unicodedata.normalize('NFKD', input_str)
    return "".join([c for c in nfkd_form if not unicodedata.combining(

def clean_name(name):
    name = name.lower()
    name = remove_accents(name)
    name = re.sub(r'[-.]', '', name)
    name = re.sub(r'\s+', ' ', name).strip()
    return name

final_2016_df['player_name'] = final_2016_df['player_name'].apply(clea
"""
```

In [290]:

```
"""
# Convert 'player_name' from "last name, first name" to "first name la
final_2016_df['Name'] = final_2016_df['player_name'].apply(lambda x: '
"""
```

In [291]:

```
"""
# Now, you can perform the merge using 'Name' and 'season' as the keys
better_2016_df = pd.merge(final_2016_df,
                          cleaning_filtered_df[['Name', 'season', 'Age
                          on=['Name', 'season'],
                          how='left')
"""
```

In [292]:

```
#better_2016_df.to_csv('data/better_2016_df.csv')
```

In [184]:
```
1 #better_2016_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1475 entries, 0 to 1474
Data columns (total 19 columns):
 #   Column                            Non-Null Count  Dtype
---  ------                            --------------  -----
 0   pitcher                           1475 non-null   int64
 1   player_name                       1475 non-null   object
 2   season                            1475 non-null   int32
 3   season_total_pitches              1475 non-null   int64
 4   pitch_type                        1475 non-null   object
 5   season_total_count_by_pitch_type  1475 non-null   int64
 6   count_by_pitch_type               1475 non-null   int64
 7   release_speed_weighted_avg        1475 non-null   float64
 8   release_pos_x_weighted_avg        1475 non-null   float64
 9   release_pos_z_weighted_avg        1475 non-null   float64
 10  vx0_weighted_avg                  1475 non-null   float64
 11  vy0_weighted_avg                  1475 non-null   float64
 12  vz0_weighted_avg                  1475 non-null   float64
 13  ax_weighted_avg                   1475 non-null   float64
 14  ay_weighted_avg                   1475 non-null   float64
 15  az_weighted_avg                   1475 non-null   float64
 16  release_pos_y_weighted_avg        1475 non-null   float64
 17  Name                              1475 non-null   object
 18  Age                               1276 non-null   float64
dtypes: float64(11), int32(1), int64(4), object(3)
memory usage: 213.3+ KB
```

In [185]:
```
1 #better_2016_df.head()
```

Out[185]:

	pitcher	player_name	season	season_total_pitches	pitch_type	season_total_count_by_pit
0	112526	colon, bartolo	2015	11167	CH	
1	112526	colon, bartolo	2015	11167	CU	
2	112526	colon, bartolo	2015	11167	FF	
3	112526	colon, bartolo	2015	11167	IN	
4	112526	colon, bartolo	2015	11167	SI	

2017

In [186]:
```
1 #all_2017_stats_df = pd.read_csv('~/Documents/Flatiron/Project_5_/data/
```

In [187]:
```
"""all_2017_stats_df.drop(columns=['batter', 'events', 'description',
                                   'des', 'game_type', 'stand', 'home_tea
                                   'away_team', 'type', 'hit_location', '
                                   'balls', 'strikes', 'pfx_x', 'spin_dir
                                   'pfx_z', 'plate_x', 'plate_z', 'on_3b'
                                   'on_2b', 'on_1b', 'outs_when_up', 'inn
                                   'inning_topbot', 'hc_x', 'hc_y', 'fiel
                                   'umpire', 'sv_id', 'hit_distance_sc',
                                    'sz_bot', 'launch_speed', 'launch_ang
                                   'pitcher.1', 'fielder_2.1', 'fielder_3
                                   'fielder_5', 'fielder_6', 'fielder_7',
                                   'fielder_9', 'estimated_ba_using_speed
                                   'estimated_woba_using_speedangle', 'ba
                                   'launch_speed_angle', 'woba_value', 'w
                                   'at_bat_number', 'pitch_number', 'home_
                                   'bat_score', 'fld_score', 'post_home_s
                                   'post_fld_score', 'post_away_score', '
                                   'of_fielding_alignment', 'delta_home_w
                                   'delta_run_exp', 'spin_rate_deprecated
                                  'break_length_deprecated', 'tfs_deprecat
all_2017_stats_df.head()
"""
```

Out[187]:

	pitch_type	game_date	release_speed	release_pos_x	release_pos_z	player_name	pitche
0	FF	2017-09-07	94.8	-2.68	6.27	Benoit, Joaquín	27654
1	CH	2017-09-07	85.1	-2.84	6.10	Benoit, Joaquín	27654
2	FF	2017-09-07	94.7	-2.77	6.31	Benoit, Joaquín	27654
3	CH	2017-09-07	85.1	-2.88	5.96	Benoit, Joaquín	27654
4	SI	2017-09-07	94.6	-2.73	6.25	Benoit, Joaquín	27654

5 rows × 21 columns

In [188]:
```
"""
all_2017_stats_df.drop(columns=['spin_axis', 'effective_speed',
                                'release_spin_rate', 'release_extension
"""
```

In [189]:
```
#all_2017_stats_df = all_2017_stats_df.dropna(axis=0)
```

In [190]:
```
#all_2017_stats_df.reset_index(inplace=True)
```

In [191]:
```
#all_2017_stats_df.drop('index', axis=1)
```

Out[191]:

	pitch_type	game_date	release_speed	release_pos_x	release_pos_z	player_name
0	FF	2017-09-07	94.8	-2.68	6.27	Benoit, Joaquín
1	CH	2017-09-07	85.1	-2.84	6.10	Benoit, Joaquín
2	FF	2017-09-07	94.7	-2.77	6.31	Benoit, Joaquín
3	CH	2017-09-07	85.1	-2.88	5.96	Benoit, Joaquín
4	SI	2017-09-07	94.6	-2.73	6.25	Benoit, Joaquín
...	...	...	...	...	...	...
413972	FF	2017-08-04	94.8	-1.18	6.08	Woodruff, Brandon
413973	CH	2017-08-04	84.9	-1.30	6.01	Woodruff, Brandon
413974	FF	2017-08-04	95.7	-1.36	6.19	Woodruff, Brandon
413975	FF	2017-08-04	94.7	-1.45	6.13	Woodruff, Brandon
413976	FF	2017-08-04	93.4	-1.29	6.24	Woodruff, Brandon

413977 rows × 17 columns

In [192]:

```
"""# Group by 'game_date' and 'pitcher' to calculate the total pitches
total_pitches = all_2017_stats_df.groupby(['game_date', 'pitcher', 'pl

# Group by 'game_date', 'pitcher', and 'pitch_type' to calculate the s
total_pitches_by_type = all_2017_stats_df.groupby(['game_date', 'pitch

# Calculate averages of the specified metrics for each pitch type, gro
avg_metrics = all_2017_stats_df.groupby(['game_date', 'pitcher', 'play
    'release_speed': 'mean',
    'release_pos_x': 'mean',
    'release_pos_z': 'mean',
    'vx0': 'mean',
    'vy0': 'mean',
    'vz0': 'mean',
    'ax': 'mean',
    'ay': 'mean',
    'az': 'mean',
    'release_pos_y': 'mean',
}).reset_index()

grouped_2017_df = total_pitches.merge(total_pitches_by_type, on=['game_
grouped_2017_df = grouped_2017_df.merge(avg_metrics, on=['game_date',

grouped_2017_df.head()
"""
```

Out[192]:

	game_date	pitcher	player_name	total_pitches	pitch_type	count_by_pitch_type	release_
0	2017-04-02	407822	De La Rosa, Jorge	7	FF	4	93.8
1	2017-04-02	407822	De La Rosa, Jorge	7	FS	3	84.5
2	2017-04-02	425844	Greinke, Zack	91	CH	17	87.2
3	2017-04-02	425844	Greinke, Zack	91	CU	9	75.1
4	2017-04-02	425844	Greinke, Zack	91	FF	7	90.9

In [193]:

```
"""
grouped_2017_df['game_date'] = pd.to_datetime(grouped_2017_df['game_da
grouped_2017_df['season'] = grouped_2017_df['game_date'].dt.year

# Step 1: Season Total Pitches
season_total_pitches = grouped_2017_df.groupby(['pitcher', 'player_nam

# Step 2: Season Total by Pitch Type
season_total_by_pitch_type = grouped_2017_df.groupby(['pitcher', 'play

# Weighted Averages Calculation Setup
weighted_avg_columns = ['release_speed', 'release_pos_x', 'release_pos_
for col in weighted_avg_columns:
    grouped_2017_df[f'{col}_product'] = grouped_2017_df[col] * grouped_

weighted_avg_aggregations = {f'{col}_product': 'sum' for col in weight
weighted_avg_aggregations['count_by_pitch_type'] = 'sum'

# Aggregate for weighted averages
weighted_avg_df = grouped_2017_df.groupby(['pitcher', 'player_name', '

# Calculate weighted averages
for col in weighted_avg_columns:
    weighted_avg_df[f'{col}_weighted_avg'] = weighted_avg_df[f'{col}_p

# Cleanup
weighted_avg_df.drop(columns=[f'{col}_product' for col in weighted_avg_

# Merge season totals and weighted averages
final_2017_df = pd.merge(season_total_pitches, season_total_by_pitch_ty
final_2017_df = pd.merge(final_2017_df, weighted_avg_df, on=['pitcher'

final_2017_df.head()
"""
```

Out[193]:

	pitcher	player_name	season	season_total_pitches	pitch_type	season_total_count_by_pit
0	112526	Colon, Bartolo	2017	9440	CH	
1	112526	Colon, Bartolo	2017	9440	FC	
2	112526	Colon, Bartolo	2017	9440	FF	
3	112526	Colon, Bartolo	2017	9440	SI	
4	112526	Colon, Bartolo	2017	9440	SL	

In [194]:

```
"""final_2017_df['player_name'] = final_2017_df['player_name'].str.lowe
final_2017_df.head()
"""
```

Out[194]:

	pitcher	player_name	season	season_total_pitches	pitch_type	season_total_count_by_pit
0	112526	colon, bartolo	2017	9440	CH	
1	112526	colon, bartolo	2017	9440	FC	
2	112526	colon, bartolo	2017	9440	FF	
3	112526	colon, bartolo	2017	9440	SI	
4	112526	colon, bartolo	2017	9440	SL	

In [195]:
```
1 #print(final_2017_df['player_name'].unique())
```

```
['colon, bartolo' 'arroyo, bronson' 'benoit, joaquín' 'belisle, matt'
 'sabathia, cc' 'dickey, r.a.' 'lackey, john' 'de la rosa, jorge'
 'pérez, oliver' 'guthrie, jeremy' 'wainwright, adam' 'greinke, zack'
 'jackson, edwin' 'santana, ervin' 'blanton, joe' 'mcgowan, dustin'
 'cain, matt' 'hamels, cole' 'young, chris' 'hernández, félix'
 'petit, yusmeiro' 'verlander, justin' 'liriano, francisco'
 'jiménez, ubaldo' 'hammel, jason' 'sánchez, aníbal' 'duke, zach'
 'mccarthy, brandon' 'feldman, scott' 'nolasco, ricky' 'chavez, jesse'
 'kluber, corey' 'cecil, brett' 'hill, rich' 'shields, james'
 'garcía, jaime' 'guerra, junior' 'harrell, lucas' 'volquez, edinson'
 'vargas, jason' 'weaver, jered' 'fister, doug' 'davis, wade'
 'gallardo, yovani' 'lester, jon' 'kendrick, kyle' 'kennedy, ian'
 'miller, andrew' 'wright, steven' 'leblanc, wade' 'scherzer, max'
 'buchholz, clay' 'morrow, brandon' 'richard, clayton' 'arrieta, jake'
 'price, david' 'griffin, a.j.' 'cueto, johnny' 'bailey, homer'
 'happ, j.a.' 'lynn, lance' 'volstad, chris' 'tomlin, josh'
 'pelfrey, mike' 'gonzález, gio' 'hughes, phil' 'swarzak, anthony'
 'estrada, marco' 'nova, iván' 'chacín, jhoulys' 'carrasco, carlos'
 'anderson, brett' 'worley, vance' 'ross, tyson' 'wood, travis'
 'phelps, david' 'hellickson, jeremy' 'kershaw, clayton' 'cashner, andre
w'
 'duensing, brian' 'hunter, tommy' 'miley, wade' 'stammen, craig'
 'garza, matt' 'quintana, jose' 'pineda, michael' 'tillman, chris'
 'minor, mike' 'latos, mat' 'norris, bud' 'archer, chris' 'gibson, kyle'
 'locke, jeff' 'mcallister, zach' 'cobb, alex' 'samardzija, jeff'
 'leake, mike' 'cahill, trevor' 'santiago, héctor' 'anderson, chase'
 'holland, derek' 'peacock, brad' 'peralta, wily' 'nicasio, juan'
 'darvish, yu' 'alvarez iii, henderson' 'delgado, randall'
 'salazar, danny' 'bettis, chad' 'bumgarner, madison' 'collmenter, josh'
 'duffy, danny' 'gee, dillon' 'harvey, matt' 'moore, matt' 'nelson, jimm
y'
 'pomeranz, drew' 'porcello, rick' 'sale, chris' 'zimmermann, jordan'
 'de la rosa, rubby' 'pérez, martín' 'teheran, julio' 'shoemaker, matt'
 'anderson, tyler' 'andriese, matt' 'chatwood, tyler' 'cole, gerrit'
 'cosart, jarred' 'gray, sonny' 'hendricks, kyle' 'hudson, daniel'
 'koehler, tom' 'lyles, jordan' 'mchugh, collin' 'milone, tommy'
 'montgomery, mike' 'odorizzi, jake' 'roark, tanner' 'strasburg, stephen'
 'bauer, trevor' 'turner, jacob' 'lorenzen, michael' 'iwakuma, hisashi'
 'tanaka, masahiro' 'ryu, hyun jin' 'stripling, ross' 'wheeler, zack'
 'ureña, josé' 'boyd, matthew' 'corbin, patrick' 'fiers, mike'
 'heaney, andrew' 'matz, steven' 'miller, shelby' 'paxton, james'
 'richards, garrett' 'skaggs, tyler' 'keuchel, dallas' 'nuño, vidal'
 'straily, dan' 'stroman, marcus' 'covey, dylan' 'foltynewicz, mike'
 'gausman, kevin' 'gray, jon' 'ray, robbie' 'sanchez, aaron'
 'smith, caleb' 'syndergaard, noah' 'taillon, jameson' 'velasquez, vince'
 'walker, taijuan' 'williams, trevor' 'germán, domingo' 'montas, frankie'
 'rodriguez, eduardo' 'degrom, jacob' 'gonzales, marco' 'eickhoff, jerad'
 'junis, jakob' 'norris, daniel' 'weaver, luke' 'pivetta, nick'
 'bundy, dylan' 'clevinger, mike' 'davies, zach' 'fulmer, michael'
 'lópez, jorge' 'musgrove, joe' 'nola, aaron' 'ross, joe' 'snell, blake'
 'wisler, matt' 'woodruff, brandon' 'perdomo, luis' 'elías, roenis'
 'rodón, carlos' 'glasnow, tyler' 'fedde, erick' 'freeland, kyle'
 'lugo, seth' 'fried, max' 'giolito, lucas' 'wacha, michael'
 'márquez, germán' 'graveman, kendall' 'stratton, chris' 'suter, brent'
 'chen, wei-yin' 'eflin, zach' 'buehler, walker' 'blackburn, paul'
 'mccullers jr., lance' 'berríos, josé' 'blach, ty' 'wood, alex'
 'senzatela, antonio' 'severino, luis' 'flexen, chris' 'lópez, reynaldo'
```

```
'maeda, kenta' 'despaigne, odrisamer' 'urías, julio' 'manaea, sean'
'kuhl, chad' 'mahle, tyler' 'brault, steven' 'godley, zack'
'alcantara, sandy' 'flaherty, jack' 'hoffman, jeff' 'montgomery, jordan'
'newcomb, sean' 'lamet, dinelson']
```

In [196]:

```
"""
def remove_accents(input_str):
    nfkd_form = unicodedata.normalize('NFKD', input_str)
    return "".join([c for c in nfkd_form if not unicodedata.combining(

def clean_name(name):
    name = name.lower()
    name = remove_accents(name)
    name = re.sub(r'[-.]', '', name)
    name = re.sub(r'\s+', ' ', name).strip()
    return name

final_2017_df['player_name'] = final_2017_df['player_name'].apply(clea
"""
```

In [197]:

```
"""
# Convert 'player_name' from "last name, first name" to "first name la
final_2017_df['Name'] = final_2017_df['player_name'].apply(lambda x: '
"""
```

In [198]:

```
"""
# Now, you can perform the merge using 'Name' and 'season' as the keys
better_2017_df = pd.merge(final_2017_df,
                          cleaning_filtered_df[['Name', 'season', 'Age
                          on=['Name', 'season'],
                          how='left')
"""
```

In [199]:

```
#better_2017_df.head()
```

Out[199]:

	pitcher	player_name	season	season_total_pitches	pitch_type	season_total_count_by_pit
0	112526	colon, bartolo	2017	9440	CH	
1	112526	colon, bartolo	2017	9440	FC	
2	112526	colon, bartolo	2017	9440	FF	
3	112526	colon, bartolo	2017	9440	SI	
4	112526	colon, bartolo	2017	9440	SL	

In [200]:
```
1 #better_2017_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1204 entries, 0 to 1203
Data columns (total 19 columns):
 #   Column                            Non-Null Count  Dtype
---  ------                            --------------  -----
 0   pitcher                           1204 non-null   int64
 1   player_name                       1204 non-null   object
 2   season                            1204 non-null   int32
 3   season_total_pitches              1204 non-null   int64
 4   pitch_type                        1204 non-null   object
 5   season_total_count_by_pitch_type  1204 non-null   int64
 6   count_by_pitch_type               1204 non-null   int64
 7   release_speed_weighted_avg        1204 non-null   float64
 8   release_pos_x_weighted_avg        1204 non-null   float64
 9   release_pos_z_weighted_avg        1204 non-null   float64
 10  vx0_weighted_avg                  1204 non-null   float64
 11  vy0_weighted_avg                  1204 non-null   float64
 12  vz0_weighted_avg                  1204 non-null   float64
 13  ax_weighted_avg                   1204 non-null   float64
 14  ay_weighted_avg                   1204 non-null   float64
 15  az_weighted_avg                   1204 non-null   float64
 16  release_pos_y_weighted_avg        1204 non-null   float64
 17  Name                              1204 non-null   object
 18  Age                               1061 non-null   float64
dtypes: float64(11), int32(1), int64(4), object(3)
memory usage: 174.1+ KB
```

In [201]:
```
1 #better_2017_df.to_csv('data/better_2017_df.csv')
```

2018

In [6]:
```
1 #all_2018_stats_df = pd.read_csv('~/Documents/Flatiron/Project_5_/data/
```

In [8]:
```
1 #all_2018_stats_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 412232 entries, 0 to 412231
Data columns (total 92 columns):
 #   Column                          Non-Null Count   Dtype  
---  ------                          --------------   -----  
 0   pitch_type                      410541 non-null  object 
 1   game_date                       412232 non-null  object 
 2   release_speed                   410620 non-null  float64
 3   release_pos_x                   410601 non-null  float64
 4   release_pos_z                   410601 non-null  float64
 5   player_name                     412232 non-null  object 
 6   batter                          412232 non-null  int64  
 7   pitcher                         412232 non-null  int64  
 8   events                          106008 non-null  object 
 9   description                     412232 non-null  object 
 10  spin_dir                        0 non-null       float64
 11  spin_rate_deprecated            0 non-null       float64
 12  break_angle_deprecated          0 non-null       float64
 13  break_length_deprecated         0 non-null       float64
 14  zone                            410601 non-null  float64
 15  des                             412232 non-null  object 
 16  game_type                       412232 non-null  object 
 17  stand                           412232 non-null  object 
 18  p_throws                        412232 non-null  object 
 19  home_team                       412232 non-null  object 
 20  away_team                       412232 non-null  object 
 21  type                            412232 non-null  object 
 22  hit_location                    93420 non-null   float64
 23  bb_type                         73252 non-null   object 
 24  balls                           412232 non-null  int64  
 25  strikes                         412232 non-null  int64  
 26  game_year                       412232 non-null  int64  
 27  pfx_x                           410598 non-null  float64
 28  pfx_z                           410598 non-null  float64
 29  plate_x                         410601 non-null  float64
 30  plate_z                         410601 non-null  float64
 31  on_3b                           33416 non-null   float64
 32  on_2b                           68056 non-null   float64
 33  on_1b                           119247 non-null  float64
 34  outs_when_up                    412232 non-null  int64  
 35  inning                          412232 non-null  int64  
 36  inning_topbot                   412232 non-null  object 
 37  hc_x                            70209 non-null   float64
 38  hc_y                            70209 non-null   float64
 39  tfs_deprecated                  0 non-null       float64
 40  tfs_zulu_deprecated             0 non-null       float64
 41  fielder_2                       412232 non-null  int64  
 42  umpire                          0 non-null       float64
 43  sv_id                           0 non-null       float64
 44  vx0                             410601 non-null  float64
 45  vy0                             410601 non-null  float64
 46  vz0                             410601 non-null  float64
 47  ax                              410601 non-null  float64
 48  ay                              410601 non-null  float64
 49  az                              410601 non-null  float64
 50  sz_top                          410601 non-null  float64
 51  sz_bot                          410601 non-null  float64
```

```
 52  hit_distance_sc                  111198 non-null  float64
 53  launch_speed                     116059 non-null  float64
 54  launch_angle                     116021 non-null  float64
 55  effective_speed                  410827 non-null  float64
 56  release_spin_rate                405223 non-null  float64
 57  release_extension                410601 non-null  float64
 58  game_pk                          412232 non-null  int64
 59  pitcher.1                        412232 non-null  int64
 60  fielder_2.1                      412232 non-null  int64
 61  fielder_3                        412232 non-null  int64
 62  fielder_4                        412232 non-null  int64
 63  fielder_5                        412232 non-null  int64
 64  fielder_6                        412232 non-null  int64
 65  fielder_7                        412232 non-null  int64
 66  fielder_8                        412232 non-null  int64
 67  fielder_9                        412232 non-null  int64
 68  release_pos_y                    410601 non-null  float64
 69  estimated_ba_using_speedangle    72117 non-null   float64
 70  estimated_woba_using_speedangle  72117 non-null   float64
 71  woba_value                       106008 non-null  float64
 72  woba_denom                       104872 non-null  float64
 73  babip_value                      106008 non-null  float64
 74  iso_value                        106008 non-null  float64
 75  launch_speed_angle               72117 non-null   float64
 76  at_bat_number                    412232 non-null  int64
 77  pitch_number                     412232 non-null  int64
 78  pitch_name                       410541 non-null  object
 79  home_score                       412232 non-null  int64
 80  away_score                       412232 non-null  int64
 81  bat_score                        412232 non-null  int64
 82  fld_score                        412232 non-null  int64
 83  post_away_score                  412232 non-null  int64
 84  post_home_score                  412232 non-null  int64
 85  post_bat_score                   412232 non-null  int64
 86  post_fld_score                   412232 non-null  int64
 87  if_fielding_alignment            410569 non-null  object
 88  of_fielding_alignment            410569 non-null  object
 89  spin_axis                        410601 non-null  float64
 90  delta_home_win_exp               412232 non-null  float64
 91  delta_run_exp                    412180 non-null  float64
dtypes: float64(47), int64(28), object(17)
memory usage: 292.5+ MB
```

In [203]:

```
"""
all_2018_stats_df.drop(columns=['batter', 'events', 'description', 'zo
                                'des', 'game_type', 'stand', 'home_tea
                                'away_team', 'type', 'hit_location', '
                                'balls', 'strikes', 'pfx_x', 'spin_dir
                                'pfx_z', 'plate_x', 'plate_z', 'on_3b'
                                'on_2b', 'on_1b', 'outs_when_up', 'inn
                                'inning_topbot', 'hc_x', 'hc_y', 'fiel
                                'umpire', 'sv_id', 'hit_distance_sc',
                                 'sz_bot', 'launch_speed', 'launch_ang
                                'pitcher.1', 'fielder_2.1', 'fielder_3
                                'fielder_5', 'fielder_6', 'fielder_7',
                                'fielder_9', 'estimated_ba_using_speed
                                'estimated_woba_using_speedangle', 'ba
                                'launch_speed_angle', 'woba_value', 'w
                                'at_bat_number', 'pitch_number', 'home_
                                'bat_score', 'fld_score', 'post_home_s
                                'post_fld_score', 'post_away_score', '
                                'of_fielding_alignment', 'delta_home_w
                                'delta_run_exp', 'spin_rate_deprecated
                               'break_length_deprecated', 'tfs_deprec
all_2018_stats_df.head()
"""
```

Out[203]:

	pitch_type	game_date	release_speed	release_pos_x	release_pos_z	player_name	pitche
0	FF	2018-09-29	93.3	2.20	5.41	Pérez, Oliver	42414
1	SI	2018-09-29	92.6	2.50	5.22	Pérez, Oliver	42414
2	SL	2018-09-29	79.2	2.72	4.81	Pérez, Oliver	42414
3	SI	2018-09-29	93.0	2.53	5.22	Pérez, Oliver	42414
4	SL	2018-09-29	77.9	2.55	5.01	Pérez, Oliver	42414

5 rows × 21 columns

In [204]:

```
"""
all_2018_stats_df.drop(columns=['spin_axis', 'effective_speed',
                                'release_spin_rate', 'release_extension
"""
```

In [205]:

```
#all_2018_stats_df = all_2018_stats_df.dropna(axis=0)
```

In [206]:

```
#all_2018_stats_df.reset_index(inplace=True)
```

In [207]:
```
1 #all_2018_stats_df.drop('index', axis=1)
```

Out[207]:

	pitch_type	game_date	release_speed	release_pos_x	release_pos_z	player_name
0	FF	2018-09-29	93.3	2.20	5.41	Pérez, Oliver
1	SI	2018-09-29	92.6	2.50	5.22	Pérez, Oliver
2	SL	2018-09-29	79.2	2.72	4.81	Pérez, Oliver
3	SI	2018-09-29	93.0	2.53	5.22	Pérez, Oliver
4	SL	2018-09-29	77.9	2.55	5.01	Pérez, Oliver
...	...	...	...	...	...	...
410536	CH	2018-03-31	81.4	2.79	5.52	Yarbrough, Ryan
410537	CH	2018-03-31	80.9	2.80	5.51	Yarbrough, Ryan
410538	FC	2018-03-31	86.4	2.97	5.41	Yarbrough, Ryan
410539	CH	2018-03-31	82.0	2.77	5.51	Yarbrough, Ryan
410540	SI	2018-03-31	88.5	2.70	5.76	Yarbrough, Ryan

410541 rows × 17 columns

In [208]:

```
"""
# Group by 'game_date' and 'pitcher' to calculate the total pitches
total_pitches = all_2018_stats_df.groupby(['game_date', 'pitcher', 'pla

# Group by 'game_date', 'pitcher', and 'pitch_type' to calculate the su
total_pitches_by_type = all_2018_stats_df.groupby(['game_date', 'pitch

# Calculate averages of the specified metrics for each pitch type, grou
avg_metrics = all_2018_stats_df.groupby(['game_date', 'pitcher', 'play
    'release_speed': 'mean',
    'release_pos_x': 'mean',
    'release_pos_z': 'mean',
    'vx0': 'mean',
    'vy0': 'mean',
    'vz0': 'mean',
    'ax': 'mean',
    'ay': 'mean',
    'az': 'mean',
    'release_pos_y': 'mean',
}).reset_index()

grouped_2018_df = total_pitches.merge(total_pitches_by_type, on=['game_
grouped_2018_df = grouped_2018_df.merge(avg_metrics, on=['game_date',

grouped_2018_df.head()
"""
```

Out[208]:

	game_date	pitcher	player_name	total_pitches	pitch_type	count_by_pitch_type	release_
0	2018-03-29	407822	De La Rosa, Jorge	4	FC	2	84.3
1	2018-03-29	407822	De La Rosa, Jorge	4	FF	2	92.1
2	2018-03-29	430935	Hamels, Cole	93	CH	14	81.8
3	2018-03-29	430935	Hamels, Cole	93	CU	22	77.8
4	2018-03-29	430935	Hamels, Cole	93	FC	22	86.5

In [209]:

```
"""grouped_2018_df['game_date'] = pd.to_datetime(grouped_2018_df['game_
grouped_2018_df['season'] = grouped_2018_df['game_date'].dt.year

# Step 1: Season Total Pitches
season_total_pitches = grouped_2018_df.groupby(['pitcher', 'player_nam

# Step 2: Season Total by Pitch Type
season_total_by_pitch_type = grouped_2018_df.groupby(['pitcher', 'play

# Weighted Averages Calculation Setup
weighted_avg_columns = ['release_speed', 'release_pos_x', 'release_pos_
for col in weighted_avg_columns:
    grouped_2018_df[f'{col}_product'] = grouped_2018_df[col] * grouped_

weighted_avg_aggregations = {f'{col}_product': 'sum' for col in weight
weighted_avg_aggregations['count_by_pitch_type'] = 'sum'

# Aggregate for weighted averages
weighted_avg_df = grouped_2018_df.groupby(['pitcher', 'player_name', '

# Calculate weighted averages
for col in weighted_avg_columns:
    weighted_avg_df[f'{col}_weighted_avg'] = weighted_avg_df[f'{col}_p

# Cleanup
weighted_avg_df.drop(columns=[f'{col}_product' for col in weighted_avg_

# Merge season totals and weighted averages
final_2018_df = pd.merge(season_total_pitches, season_total_by_pitch_ty
final_2018_df = pd.merge(final_2018_df, weighted_avg_df, on=['pitcher'

final_2018_df.head()
"""
```

Out[209]:

	pitcher	player_name	season	season_total_pitches	pitch_type	season_total_count_by_pit
0	112526	Colon, Bartolo	2018	10240	CH	
1	112526	Colon, Bartolo	2018	10240	FC	
2	112526	Colon, Bartolo	2018	10240	FF	
3	112526	Colon, Bartolo	2018	10240	SI	
4	112526	Colon, Bartolo	2018	10240	SL	

In [210]:
```
"""
final_2018_df['player_name'] = final_2018_df['player_name'].str.lower(
final_2018_df.head()
"""
```

Out[210]:

	pitcher	player_name	season	season_total_pitches	pitch_type	season_total_count_by_pit
0	112526	colon, bartolo	2018	10240	CH	
1	112526	colon, bartolo	2018	10240	FC	
2	112526	colon, bartolo	2018	10240	FF	
3	112526	colon, bartolo	2018	10240	SI	
4	112526	colon, bartolo	2018	10240	SL	

In [211]:
```
"""
def remove_accents(input_str):
    nfkd_form = unicodedata.normalize('NFKD', input_str)
    return "".join([c for c in nfkd_form if not unicodedata.combining(

def clean_name(name):
    name = name.lower()
    name = remove_accents(name)
    name = re.sub(r'[-.]', '', name)
    name = re.sub(r'\s+', ' ', name).strip()
    return name

final_2018_df['player_name'] = final_2018_df['player_name'].apply(clea
"""
```

In [212]:
```
"""
# Convert 'player_name' from "last name, first name" to "first name la
final_2018_df['Name'] = final_2018_df['player_name'].apply(lambda x: '
"""
```

In [213]:
```
"""
# Now, you can perform the merge using 'Name' and 'season' as the keys
better_2018_df = pd.merge(final_2018_df,
                          cleaning_filtered_df[['Name', 'season', 'Age
                          on=['Name', 'season'],
                          how='left')
"""
```

In [214]:
```
#better_2018_df.head()
```

Out[214]:

	pitcher	player_name	season	season_total_pitches	pitch_type	season_total_count_by_pit
0	112526	colon, bartolo	2018	10240	CH	
1	112526	colon, bartolo	2018	10240	FC	
2	112526	colon, bartolo	2018	10240	FF	
3	112526	colon, bartolo	2018	10240	SI	
4	112526	colon, bartolo	2018	10240	SL	

In [215]:
```
#better_2018_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1192 entries, 0 to 1191
Data columns (total 19 columns):
 #   Column                            Non-Null Count  Dtype
---  ------                            --------------  -----
 0   pitcher                           1192 non-null   int64
 1   player_name                       1192 non-null   object
 2   season                            1192 non-null   int32
 3   season_total_pitches              1192 non-null   int64
 4   pitch_type                        1192 non-null   object
 5   season_total_count_by_pitch_type  1192 non-null   int64
 6   count_by_pitch_type               1192 non-null   int64
 7   release_speed_weighted_avg        1192 non-null   float64
 8   release_pos_x_weighted_avg        1192 non-null   float64
 9   release_pos_z_weighted_avg        1192 non-null   float64
 10  vx0_weighted_avg                  1192 non-null   float64
 11  vy0_weighted_avg                  1192 non-null   float64
 12  vz0_weighted_avg                  1192 non-null   float64
 13  ax_weighted_avg                   1192 non-null   float64
 14  ay_weighted_avg                   1192 non-null   float64
 15  az_weighted_avg                   1192 non-null   float64
 16  release_pos_y_weighted_avg        1192 non-null   float64
 17  Name                              1192 non-null   object
 18  Age                               1068 non-null   float64
dtypes: float64(11), int32(1), int64(4), object(3)
memory usage: 172.4+ KB
```

In [216]:
```
#better_2018_df.to_csv('data/better_2018_df.csv')
```

2019

In [217]:
```
#all_2019_stats_df = pd.read_csv('~/Documents/Flatiron/Project_5_/data/
```

In [218]:
```
"""
all_2019_stats_df.drop(columns=['batter', 'events', 'description', 'zo
                                'des', 'game_type', 'stand', 'home_tea
                                'away_team', 'type', 'hit_location', 'l
                                'balls', 'strikes', 'pfx_x', 'spin_dir
                                'pfx_z', 'plate_x', 'plate_z', 'on_3b'
                                'on_2b', 'on_1b', 'outs_when_up', 'inn
                                'inning_topbot', 'hc_x', 'hc_y', 'fiel
                                'umpire', 'sv_id', 'hit_distance_sc',
                                 'sz_bot', 'launch_speed', 'launch_ang
                                'pitcher.1', 'fielder_2.1', 'fielder_3
                                'fielder_5', 'fielder_6', 'fielder_7',
                                'fielder_9', 'estimated_ba_using_speed
                                'estimated_woba_using_speedangle', 'bal
                                'launch_speed_angle', 'woba_value', 'w
                                'at_bat_number', 'pitch_number', 'home_
                                'bat_score', 'fld_score', 'post_home_s
                                'post_fld_score', 'post_away_score', '
                                'of_fielding_alignment', 'delta_home_w
                                'delta_run_exp', 'spin_rate_deprecated
                               'break_length_deprecated', 'tfs_deprec
all_2019_stats_df.head()
"""
```

Out[218]:

	pitch_type	game_date	release_speed	release_pos_x	release_pos_z	player_name	pitche
0	SI	2019-09-25	92.8	2.67	5.27	Pérez, Oliver	42414
1	SL	2019-09-25	79.0	2.76	5.05	Pérez, Oliver	42414
2	SI	2019-09-25	92.3	2.69	5.23	Pérez, Oliver	42414
3	SI	2019-09-25	93.4	2.50	5.46	Pérez, Oliver	42414
4	SI	2019-09-25	92.0	2.68	5.32	Pérez, Oliver	42414

5 rows × 21 columns

In [219]:
```
"""
all_2019_stats_df.drop(columns=['spin_axis', 'effective_speed',
                                'release_spin_rate', 'release_extension
"""
```

In [220]:
```
#all_2019_stats_df = all_2019_stats_df.dropna(axis=0)
```

In [221]:
```
#all_2019_stats_df.reset_index(inplace=True)
```

In [222]:
```
#all_2019_stats_df.drop('index', axis=1)
```

Out[222]:

	pitch_type	game_date	release_speed	release_pos_x	release_pos_z	player_name
0	SI	2019-09-25	92.8	2.67	5.27	Pérez, Oliver
1	SL	2019-09-25	79.0	2.76	5.05	Pérez, Oliver
2	SI	2019-09-25	92.3	2.69	5.23	Pérez, Oliver
3	SI	2019-09-25	93.4	2.50	5.46	Pérez, Oliver
4	SI	2019-09-25	92.0	2.68	5.32	Pérez, Oliver
...	...	...	...	...	...	...
401447	SI	2019-03-26	88.3	2.79	5.64	Yarbrough, Ryan
401448	FC	2019-03-26	84.3	3.08	5.19	Yarbrough, Ryan
401449	FC	2019-03-26	83.3	3.07	5.21	Yarbrough, Ryan
401450	SI	2019-03-26	86.9	2.84	5.74	Yarbrough, Ryan
401451	SI	2019-03-26	86.5	2.74	5.70	Yarbrough, Ryan

401452 rows × 17 columns

In [223]:

```
"""
# Group by 'game_date' and 'pitcher' to calculate the total pitches
total_pitches = all_2019_stats_df.groupby(['game_date', 'pitcher', 'pla

# Group by 'game_date', 'pitcher', and 'pitch_type' to calculate the su
total_pitches_by_type = all_2019_stats_df.groupby(['game_date', 'pitch

# Calculate averages of the specified metrics for each pitch type, grou
avg_metrics = all_2019_stats_df.groupby(['game_date', 'pitcher', 'play
    'release_speed': 'mean',
    'release_pos_x': 'mean',
    'release_pos_z': 'mean',
    'vx0': 'mean',
    'vy0': 'mean',
    'vz0': 'mean',
    'ax': 'mean',
    'ay': 'mean',
    'az': 'mean',
    'release_pos_y': 'mean',
}).reset_index()

grouped_2019_df = total_pitches.merge(total_pitches_by_type, on=['game_
grouped_2019_df = grouped_2019_df.merge(avg_metrics, on=['game_date',

grouped_2019_df.head()
"""
```

Out[223]:

	game_date	pitcher	player_name	total_pitches	pitch_type	count_by_pitch_type	release_
0	2019-03-20	453178	Kennedy, Ian	19	CH	1	87.4
1	2019-03-20	453178	Kennedy, Ian	19	FF	13	93.0
2	2019-03-20	453178	Kennedy, Ian	19	KC	4	79.2
3	2019-03-20	453178	Kennedy, Ian	19	SL	1	90.2
4	2019-03-20	542881	Anderson, Tyler	84	CH	24	79.8

In [224]:

```
"""
grouped_2019_df['game_date'] = pd.to_datetime(grouped_2019_df['game_da
grouped_2019_df['season'] = grouped_2019_df['game_date'].dt.year

# Step 1: Season Total Pitches
season_total_pitches = grouped_2019_df.groupby(['pitcher', 'player_nam

# Step 2: Season Total by Pitch Type
season_total_by_pitch_type = grouped_2019_df.groupby(['pitcher', 'play

# Weighted Averages Calculation Setup
weighted_avg_columns = ['release_speed', 'release_pos_x', 'release_pos_
for col in weighted_avg_columns:
    grouped_2019_df[f'{col}_product'] = grouped_2019_df[col] * grouped_

weighted_avg_aggregations = {f'{col}_product': 'sum' for col in weight
weighted_avg_aggregations['count_by_pitch_type'] = 'sum'

# Aggregate for weighted averages
weighted_avg_df = grouped_2019_df.groupby(['pitcher', 'player_name', '

# Calculate weighted averages
for col in weighted_avg_columns:
    weighted_avg_df[f'{col}_weighted_avg'] = weighted_avg_df[f'{col}_p

# Cleanup
weighted_avg_df.drop(columns=[f'{col}_product' for col in weighted_avg_

# Merge season totals and weighted averages
final_2019_df = pd.merge(season_total_pitches, season_total_by_pitch_ty
final_2019_df = pd.merge(final_2019_df, weighted_avg_df, on=['pitcher'

final_2019_df.head()
"""
```

Out[224]:

	pitcher	player_name	season	season_total_pitches	pitch_type	season_total_count_by_pit
0	282332	Sabathia, CC	2019	7786	CH	
1	282332	Sabathia, CC	2019	7786	FC	
2	282332	Sabathia, CC	2019	7786	FF	
3	282332	Sabathia, CC	2019	7786	SI	
4	282332	Sabathia, CC	2019	7786	SL	

In [225]:
```
"""
final_2019_df['player_name'] = final_2019_df['player_name'].str.lower(
final_2019_df.head()
"""
```

Out[225]:

	pitcher	player_name	season	season_total_pitches	pitch_type	season_total_count_by_pit
0	282332	sabathia, cc	2019	7786	CH	
1	282332	sabathia, cc	2019	7786	FC	
2	282332	sabathia, cc	2019	7786	FF	
3	282332	sabathia, cc	2019	7786	SI	
4	282332	sabathia, cc	2019	7786	SL	

In [226]:
```
"""
def remove_accents(input_str):
    nfkd_form = unicodedata.normalize('NFKD', input_str)
    return "".join([c for c in nfkd_form if not unicodedata.combining(

def clean_name(name):
    name = name.lower()
    name = remove_accents(name)
    name = re.sub(r'[-.]', '', name)
    name = re.sub(r'\s+', ' ', name).strip()
    return name

final_2019_df['player_name'] = final_2019_df['player_name'].apply(clea
"""
```

In [227]:
```
"""
# Convert 'player_name' from "last name, first name" to "first name la
final_2019_df['Name'] = final_2019_df['player_name'].apply(lambda x: '
"""
```

In [228]:
```
"""
# Now, you can perform the merge using 'Name' and 'season' as the keys
better_2019_df = pd.merge(final_2019_df,
                          cleaning_filtered_df[['Name', 'season', 'Age
                          on=['Name', 'season'],
                          how='left')
"""
```

In [229]:
```
#better_2019_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1154 entries, 0 to 1153
Data columns (total 19 columns):
 #   Column                            Non-Null Count  Dtype  
---  ------                            --------------  -----  
 0   pitcher                           1154 non-null   int64  
 1   player_name                       1154 non-null   object 
 2   season                            1154 non-null   int32  
 3   season_total_pitches              1154 non-null   int64  
 4   pitch_type                        1154 non-null   object 
 5   season_total_count_by_pitch_type  1154 non-null   int64  
 6   count_by_pitch_type               1154 non-null   int64  
 7   release_speed_weighted_avg        1154 non-null   float64
 8   release_pos_x_weighted_avg        1154 non-null   float64
 9   release_pos_z_weighted_avg        1154 non-null   float64
 10  vx0_weighted_avg                  1154 non-null   float64
 11  vy0_weighted_avg                  1154 non-null   float64
 12  vz0_weighted_avg                  1154 non-null   float64
 13  ax_weighted_avg                   1154 non-null   float64
 14  ay_weighted_avg                   1154 non-null   float64
 15  az_weighted_avg                   1154 non-null   float64
 16  release_pos_y_weighted_avg        1154 non-null   float64
 17  Name                              1154 non-null   object 
 18  Age                               1057 non-null   float64
dtypes: float64(11), int32(1), int64(4), object(3)
memory usage: 166.9+ KB
```

In [230]:
```
#better_2019_df.head()
```

Out[230]:

	pitcher	player_name	season	season_total_pitches	pitch_type	season_total_count_by_pit
0	282332	sabathia, cc	2019	7786	CH	
1	282332	sabathia, cc	2019	7786	FC	
2	282332	sabathia, cc	2019	7786	FF	
3	282332	sabathia, cc	2019	7786	SI	
4	282332	sabathia, cc	2019	7786	SL	

In [231]:
```
#better_2019_df.to_csv('data/better_2019_df.csv')
```

2020

In [232]:
```
#all_2020_stats_df = pd.read_csv('~/Documents/Flatiron/Project_5_/data/
```

In [233]:
```
"""all_2020_stats_df.drop(columns=['batter', 'events', 'description',
                                   'des', 'game_type', 'stand', 'home_tea
                                   'away_team', 'type', 'hit_location', '
                                   'balls', 'strikes', 'pfx_x', 'spin_dir
                                   'pfx_z', 'plate_x', 'plate_z', 'on_3b'
                                   'on_2b', 'on_1b', 'outs_when_up', 'inn
                                   'inning_topbot', 'hc_x', 'hc_y', 'fiel
                                   'umpire', 'sv_id', 'hit_distance_sc',
                                    'sz_bot', 'launch_speed', 'launch_ang
                                   'pitcher.1', 'fielder_2.1', 'fielder_3
                                   'fielder_5', 'fielder_6', 'fielder_7',
                                   'fielder_9', 'estimated_ba_using_speed
                                   'estimated_woba_using_speedangle', 'ba
                                   'launch_speed_angle', 'woba_value', 'w
                                   'at_bat_number', 'pitch_number', 'home_
                                   'bat_score', 'fld_score', 'post_home_s
                                   'post_fld_score', 'post_away_score', '
                                   'of_fielding_alignment', 'delta_home_w
                                   'delta_run_exp', 'spin_rate_deprecated
                                  'break_length_deprecated', 'tfs_deprecat
all_2020_stats_df.head()
"""
```

Out[233]:

	pitch_type	game_date	release_speed	release_pos_x	release_pos_z	player_name	pitche
0	SI	2020-09-27	88.9	2.98	5.29	Pérez, Oliver	42414
1	SL	2020-09-27	74.8	2.10	5.73	Pérez, Oliver	42414
2	SL	2020-09-27	75.3	2.01	5.74	Pérez, Oliver	42414
3	SL	2020-09-27	75.8	2.38	5.20	Pérez, Oliver	42414
4	SL	2020-09-27	75.0	2.70	5.15	Pérez, Oliver	42414

5 rows × 21 columns

In [234]:
```
"""
all_2020_stats_df.drop(columns=['spin_axis', 'effective_speed',
                                'release_spin_rate', 'release_extension
"""
```

In [235]:
```
#all_2020_stats_df = all_2020_stats_df.dropna(axis=0)
```

In [236]:
```
#all_2020_stats_df.reset_index(inplace=True)
```

In [237]:
```
1 #all_2020_stats_df.drop('index', axis=1)
```

Out[237]:

	pitch_type	game_date	release_speed	release_pos_x	release_pos_z	player_name
0	SI	2020-09-27	88.9	2.98	5.29	Pérez, Oliver
1	SL	2020-09-27	74.8	2.10	5.73	Pérez, Oliver
2	SL	2020-09-27	75.3	2.01	5.74	Pérez, Oliver
3	SL	2020-09-27	75.8	2.38	5.20	Pérez, Oliver
4	SL	2020-09-27	75.0	2.70	5.15	Pérez, Oliver
...	...	...	...	...	...	...
131170	FC	2020-07-25	84.5	2.99	4.99	Yarbrough, Ryan
131171	FC	2020-07-25	86.2	2.93	4.93	Yarbrough, Ryan
131172	FC	2020-07-25	84.0	3.08	4.95	Yarbrough, Ryan
131173	FC	2020-07-25	83.3	2.91	5.03	Yarbrough, Ryan
131174	SI	2020-07-25	88.3	2.50	5.62	Yarbrough, Ryan

131175 rows × 17 columns

In [238]:

```
"""
# Group by 'game_date' and 'pitcher' to calculate the total pitches
total_pitches = all_2020_stats_df.groupby(['game_date', 'pitcher', 'pla

# Group by 'game_date', 'pitcher', and 'pitch_type' to calculate the su
total_pitches_by_type = all_2020_stats_df.groupby(['game_date', 'pitch

# Calculate averages of the specified metrics for each pitch type, grou
avg_metrics = all_2020_stats_df.groupby(['game_date', 'pitcher', 'play
    'release_speed': 'mean',
    'release_pos_x': 'mean',
    'release_pos_z': 'mean',
    'vx0': 'mean',
    'vy0': 'mean',
    'vz0': 'mean',
    'ax': 'mean',
    'ay': 'mean',
    'az': 'mean',
    'release_pos_y': 'mean',
}).reset_index()

grouped_2020_df = total_pitches.merge(total_pitches_by_type, on=['game_
grouped_2020_df = grouped_2020_df.merge(avg_metrics, on=['game_date',

grouped_2020_df.head()
"""
```

Out[238]:

	game_date	pitcher	player_name	total_pitches	pitch_type	count_by_pitch_type	release_
0	2020-07-23	453286	Scherzer, Max	99	CH	10	85.1
1	2020-07-23	453286	Scherzer, Max	99	CU	8	76.9
2	2020-07-23	453286	Scherzer, Max	99	FC	4	91.6
3	2020-07-23	453286	Scherzer, Max	99	FF	43	95.0
4	2020-07-23	453286	Scherzer, Max	99	SL	34	86.2

In [239]:

```
"""
grouped_2020_df['game_date'] = pd.to_datetime(grouped_2020_df['game_da
grouped_2020_df['season'] = grouped_2020_df['game_date'].dt.year

# Step 1: Season Total Pitches
season_total_pitches = grouped_2020_df.groupby(['pitcher', 'player_nam

# Step 2: Season Total by Pitch Type
season_total_by_pitch_type = grouped_2020_df.groupby(['pitcher', 'play

# Weighted Averages Calculation Setup
weighted_avg_columns = ['release_speed', 'release_pos_x', 'release_pos_
for col in weighted_avg_columns:
    grouped_2020_df[f'{col}_product'] = grouped_2020_df[col] * grouped_

weighted_avg_aggregations = {f'{col}_product': 'sum' for col in weight
weighted_avg_aggregations['count_by_pitch_type'] = 'sum'

# Aggregate for weighted averages
weighted_avg_df = grouped_2020_df.groupby(['pitcher', 'player_name', '

# Calculate weighted averages
for col in weighted_avg_columns:
    weighted_avg_df[f'{col}_weighted_avg'] = weighted_avg_df[f'{col}_p

# Cleanup
weighted_avg_df.drop(columns=[f'{col}_product' for col in weighted_avg_

# Merge season totals and weighted averages
final_2020_df = pd.merge(season_total_pitches, season_total_by_pitch_ty
final_2020_df = pd.merge(final_2020_df, weighted_avg_df, on=['pitcher'

final_2020_df.head()
"""
```

Out[239]:

	pitcher	player_name	season	season_total_pitches	pitch_type	season_total_count_by_pit
0	424144	Pérez, Oliver	2020	810	FF	
1	424144	Pérez, Oliver	2020	810	SI	
2	424144	Pérez, Oliver	2020	810	SL	
3	425794	Wainwright, Adam	2020	4559	CH	
4	425794	Wainwright, Adam	2020	4559	CU	

In [240]:
```
"""
final_2020_df['player_name'] = final_2020_df['player_name'].str.lower(
final_2020_df.head()
"""
```

Out[240]:

	pitcher	player_name	season	season_total_pitches	pitch_type	season_total_count_by_pit
0	424144	pérez, oliver	2020	810	FF	
1	424144	pérez, oliver	2020	810	SI	
2	424144	pérez, oliver	2020	810	SL	
3	425794	wainwright, adam	2020	4559	CH	
4	425794	wainwright, adam	2020	4559	CU	

In [241]:
```
"""
def remove_accents(input_str):
    nfkd_form = unicodedata.normalize('NFKD', input_str)
    return "".join([c for c in nfkd_form if not unicodedata.combining(
def clean_name(name):
    name = name.lower()
    name = remove_accents(name)
    name = re.sub(r'[-.]', '', name)
    name = re.sub(r'\s+', ' ', name).strip()
    return name
final_2020_df['player_name'] = final_2020_df['player_name'].apply(clea
"""
```

In [242]:
```
"""
# Convert 'player_name' from "last name, first name" to "first name la
final_2020_df['Name'] = final_2020_df['player_name'].apply(lambda x: '
"""
```

In [243]:
```
"""# Now, you can perform the merge using 'Name' and 'season' as the k
better_2020_df = pd.merge(final_2020_df,
                          cleaning_filtered_df[['Name', 'season', 'Age
                          on=['Name', 'season'],
                          how='left')
"""
```

In [244]:
```
#better_2020_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 945 entries, 0 to 944
Data columns (total 19 columns):
 #   Column                            Non-Null Count  Dtype
---  ------                            --------------  -----
 0   pitcher                           945 non-null    int64
 1   player_name                       945 non-null    object
 2   season                            945 non-null    int32
 3   season_total_pitches              945 non-null    int64
 4   pitch_type                        945 non-null    object
 5   season_total_count_by_pitch_type  945 non-null    int64
 6   count_by_pitch_type               945 non-null    int64
 7   release_speed_weighted_avg        945 non-null    float64
 8   release_pos_x_weighted_avg        945 non-null    float64
 9   release_pos_z_weighted_avg        945 non-null    float64
 10  vx0_weighted_avg                  945 non-null    float64
 11  vy0_weighted_avg                  945 non-null    float64
 12  vz0_weighted_avg                  945 non-null    float64
 13  ax_weighted_avg                   945 non-null    float64
 14  ay_weighted_avg                   945 non-null    float64
 15  az_weighted_avg                   945 non-null    float64
 16  release_pos_y_weighted_avg        945 non-null    float64
 17  Name                              945 non-null    object
 18  Age                               856 non-null    float64
dtypes: float64(11), int32(1), int64(4), object(3)
memory usage: 136.7+ KB
```

In [245]:
```
#better_2020_df.head()
```

Out[245]:

	pitcher	player_name	season	season_total_pitches	pitch_type	season_total_count_by_pit
0	424144	perez, oliver	2020	810	FF	
1	424144	perez, oliver	2020	810	SI	
2	424144	perez, oliver	2020	810	SL	
3	425794	wainwright, adam	2020	4559	CH	
4	425794	wainwright, adam	2020	4559	CU	

In [246]:
```
#better_2020_df.to_csv('data/better_2020_df.csv')
```

2021

In [247]:
```
#all_2021_stats_df = pd.read_csv('~/Documents/Flatiron/Project_5_/data/
```

In [248]:
```
"""
all_2021_stats_df.drop(columns=['batter', 'events', 'description', 'zo
                                'des', 'game_type', 'stand', 'home_tea
                                'away_team', 'type', 'hit_location', 'l
                                'balls', 'strikes', 'pfx_x', 'spin_dir
                                'pfx_z', 'plate_x', 'plate_z', 'on_3b'
                                'on_2b', 'on_1b', 'outs_when_up', 'inn
                                'inning_topbot', 'hc_x', 'hc_y', 'fiel
                                'umpire', 'sv_id', 'hit_distance_sc',
                                 'sz_bot', 'launch_speed', 'launch_ang
                                'pitcher.1', 'fielder_2.1', 'fielder_3
                                'fielder_5', 'fielder_6', 'fielder_7',
                                'fielder_9', 'estimated_ba_using_speed
                                'estimated_woba_using_speedangle', 'bal
                                'launch_speed_angle', 'woba_value', 'w
                                'at_bat_number', 'pitch_number', 'home_
                                'bat_score', 'fld_score', 'post_home_s
                                'post_fld_score', 'post_away_score', '
                                'of_fielding_alignment', 'delta_home_w
                                'delta_run_exp', 'spin_rate_deprecated
                               'break_length_deprecated', 'tfs_deprec
all_2021_stats_df.head()
"""
```

Out[248]:

	pitch_type	game_date	release_speed	release_pos_x	release_pos_z	player_name	pitche
0	SL	2021-04-22	77.1	2.64	5.04	Pérez, Oliver	42414
1	SI	2021-04-22	89.2	2.40	5.35	Pérez, Oliver	42414
2	SL	2021-04-22	75.8	2.80	5.12	Pérez, Oliver	42414
3	SL	2021-04-22	75.7	2.49	5.24	Pérez, Oliver	42414
4	SL	2021-04-22	77.5	2.51	5.35	Pérez, Oliver	42414

5 rows × 21 columns

In [249]:
```
"""
all_2021_stats_df.drop(columns=['spin_axis', 'effective_speed',
                                'release_spin_rate', 'release_extension
"""
```

In [250]:
```
#all_2021_stats_df = all_2021_stats_df.dropna(axis=0)
```

In [251]:
```
#all_2021_stats_df.reset_index(inplace=True)
```

In [252]:
```
1 #all_2021_stats_df.drop('index', axis=1)
```

Out[252]:

	pitch_type	game_date	release_speed	release_pos_x	release_pos_z	player_name
0	SL	2021-04-22	77.1	2.64	5.04	Pérez, Oliver
1	SI	2021-04-22	89.2	2.40	5.35	Pérez, Oliver
2	SL	2021-04-22	75.8	2.80	5.12	Pérez, Oliver
3	SL	2021-04-22	75.7	2.49	5.24	Pérez, Oliver
4	SL	2021-04-22	77.5	2.51	5.35	Pérez, Oliver
...	...	...	...	...	...	...
349392	FC	2021-04-02	80.4	2.97	5.19	Yarbrough, Ryan
349393	FC	2021-04-02	78.0	2.93	5.33	Yarbrough, Ryan
349394	CH	2021-04-02	77.7	2.40	5.72	Yarbrough, Ryan
349395	CH	2021-04-02	78.3	2.41	5.66	Yarbrough, Ryan
349396	SI	2021-04-02	84.0	2.32	5.97	Yarbrough, Ryan

349397 rows × 17 columns

In [253]:

```
"""# Group by 'game_date' and 'pitcher' to calculate the total pitches
total_pitches = all_2021_stats_df.groupby(['game_date', 'pitcher', 'pl

# Group by 'game_date', 'pitcher', and 'pitch_type' to calculate the s
total_pitches_by_type = all_2021_stats_df.groupby(['game_date', 'pitch

# Calculate averages of the specified metrics for each pitch type, gro
avg_metrics = all_2021_stats_df.groupby(['game_date', 'pitcher', 'play
    'release_speed': 'mean',
    'release_pos_x': 'mean',
    'release_pos_z': 'mean',
    'vx0': 'mean',
    'vy0': 'mean',
    'vz0': 'mean',
    'ax': 'mean',
    'ay': 'mean',
    'az': 'mean',
    'release_pos_y': 'mean',
}).reset_index()

grouped_2021_df = total_pitches.merge(total_pitches_by_type, on=['game_
grouped_2021_df = grouped_2021_df.merge(avg_metrics, on=['game_date',

grouped_2021_df.head()
"""
```

Out[253]:

	game_date	pitcher	player_name	total_pitches	pitch_type	count_by_pitch_type	release_
0	2021-04-01	425844	Greinke, Zack	82	CH	20	86.9
1	2021-04-01	425844	Greinke, Zack	82	CU	14	71.9
2	2021-04-01	425844	Greinke, Zack	82	FF	35	88.6
3	2021-04-01	425844	Greinke, Zack	82	SL	13	81.5
4	2021-04-01	433589	Petit, Yusmeiro	11	CH	2	79.7

In [254]:

```
"""
grouped_2021_df['game_date'] = pd.to_datetime(grouped_2021_df['game_da
grouped_2021_df['season'] = grouped_2021_df['game_date'].dt.year

# Step 1: Season Total Pitches
season_total_pitches = grouped_2021_df.groupby(['pitcher', 'player_nam

# Step 2: Season Total by Pitch Type
season_total_by_pitch_type = grouped_2021_df.groupby(['pitcher', 'play

# Weighted Averages Calculation Setup
weighted_avg_columns = ['release_speed', 'release_pos_x', 'release_pos_
for col in weighted_avg_columns:
    grouped_2021_df[f'{col}_product'] = grouped_2021_df[col] * grouped_

weighted_avg_aggregations = {f'{col}_product': 'sum' for col in weight
weighted_avg_aggregations['count_by_pitch_type'] = 'sum'

# Aggregate for weighted averages
weighted_avg_df = grouped_2021_df.groupby(['pitcher', 'player_name', '

# Calculate weighted averages
for col in weighted_avg_columns:
    weighted_avg_df[f'{col}_weighted_avg'] = weighted_avg_df[f'{col}_p

# Cleanup
weighted_avg_df.drop(columns=[f'{col}_product' for col in weighted_avg_

# Merge season totals and weighted averages
final_2021_df = pd.merge(season_total_pitches, season_total_by_pitch_ty
final_2021_df = pd.merge(final_2021_df, weighted_avg_df, on=['pitcher'

final_2021_df.head()
"""
```

Out[254]:

	pitcher	player_name	season	season_total_pitches	pitch_type	season_total_count_by_pit
0	424144	Pérez, Oliver	2021	168	FF	
1	424144	Pérez, Oliver	2021	168	SI	
2	424144	Pérez, Oliver	2021	168	SL	
3	425794	Wainwright, Adam	2021	16523	CH	
4	425794	Wainwright, Adam	2021	16523	CS	

In [255]:
```
"""
final_2021_df['player_name'] = final_2021_df['player_name'].str.lower(
final_2021_df.head()
"""
```

Out[255]:

	pitcher	player_name	season	season_total_pitches	pitch_type	season_total_count_by_pit
0	424144	pérez, oliver	2021	168	FF	
1	424144	pérez, oliver	2021	168	SI	
2	424144	pérez, oliver	2021	168	SL	
3	425794	wainwright, adam	2021	16523	CH	
4	425794	wainwright, adam	2021	16523	CS	

In [256]:
```
"""
def remove_accents(input_str):
    nfkd_form = unicodedata.normalize('NFKD', input_str)
    return "".join([c for c in nfkd_form if not unicodedata.combining(

def clean_name(name):
    name = name.lower()
    name = remove_accents(name)
    name = re.sub(r'[-.]', '', name)
    name = re.sub(r'\s+', ' ', name).strip()
    return name

final_2021_df['player_name'] = final_2021_df['player_name'].apply(clea
"""
```

In [257]:
```
"""
# Convert 'player_name' from "last name, first name" to "first name la
final_2021_df['Name'] = final_2021_df['player_name'].apply(lambda x: '
"""
```

In [258]:
```
"""
# Now, you can perform the merge using 'Name' and 'season' as the keys
better_2021_df = pd.merge(final_2021_df,
                          cleaning_filtered_df[['Name', 'season', 'Age
                          on=['Name', 'season'],
                          how='left')
"""
```

In [259]:
```
#better_2021_df.head()
```

Out[259]:

	pitcher	player_name	season	season_total_pitches	pitch_type	season_total_count_by_pit
0	424144	perez, oliver	2021	168	FF	
1	424144	perez, oliver	2021	168	SI	
2	424144	perez, oliver	2021	168	SL	
3	425794	wainwright, adam	2021	16523	CH	
4	425794	wainwright, adam	2021	16523	CS	

In [260]:
```
#better_2021_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1024 entries, 0 to 1023
Data columns (total 19 columns):
 #   Column                             Non-Null Count  Dtype  
---  ------                             --------------  -----  
 0   pitcher                            1024 non-null   int64  
 1   player_name                        1024 non-null   object 
 2   season                             1024 non-null   int32  
 3   season_total_pitches               1024 non-null   int64  
 4   pitch_type                         1024 non-null   object 
 5   season_total_count_by_pitch_type   1024 non-null   int64  
 6   count_by_pitch_type                1024 non-null   int64  
 7   release_speed_weighted_avg         1024 non-null   float64
 8   release_pos_x_weighted_avg         1024 non-null   float64
 9   release_pos_z_weighted_avg         1024 non-null   float64
 10  vx0_weighted_avg                   1024 non-null   float64
 11  vy0_weighted_avg                   1024 non-null   float64
 12  vz0_weighted_avg                   1024 non-null   float64
 13  ax_weighted_avg                    1024 non-null   float64
 14  ay_weighted_avg                    1024 non-null   float64
 15  az_weighted_avg                    1024 non-null   float64
 16  release_pos_y_weighted_avg         1024 non-null   float64
 17  Name                               1024 non-null   object 
 18  Age                                932 non-null    float64
dtypes: float64(11), int32(1), int64(4), object(3)
memory usage: 148.1+ KB
```

In [261]:
```
#better_2021_df.to_csv('data/better_2021_df.csv')
```

2022

In [262]:
```
#all_2022_stats_df = pd.read_csv('~/Documents/Flatiron/Project_5_/data/
```

In [263]:
```
"""all_2022_stats_df.drop(columns=['batter', 'events', 'description',
                                   'des', 'game_type', 'stand', 'home_tea
                                   'away_team', 'type', 'hit_location', 'l
                                   'balls', 'strikes', 'pfx_x', 'spin_dir
                                   'pfx_z', 'plate_x', 'plate_z', 'on_3b'
                                   'on_2b', 'on_1b', 'outs_when_up', 'inn
                                   'inning_topbot', 'hc_x', 'hc_y', 'fiel
                                   'umpire', 'sv_id', 'hit_distance_sc',
                                    'sz_bot', 'launch_speed', 'launch_ang
                                   'pitcher.1', 'fielder_2.1', 'fielder_3
                                   'fielder_5', 'fielder_6', 'fielder_7',
                                   'fielder_9', 'estimated_ba_using_speed
                                   'estimated_woba_using_speedangle', 'bal
                                   'launch_speed_angle', 'woba_value', 'w
                                   'at_bat_number', 'pitch_number', 'home_
                                   'bat_score', 'fld_score', 'post_home_s
                                   'post_fld_score', 'post_away_score', '
                                   'of_fielding_alignment', 'delta_home_w
                                   'delta_run_exp', 'spin_rate_deprecated
                                  'break_length_deprecated', 'tfs_deprecat
all_2022_stats_df.head()
"""
```

Out[263]:

	pitch_type	game_date	release_speed	release_pos_x	release_pos_z	player_name	pitche
0	SI	2022-04-24	87.8	2.74	5.20	Pérez, Oliver	42414
1	SL	2022-04-24	75.8	2.12	5.60	Pérez, Oliver	42414
2	SI	2022-04-24	87.3	2.45	5.41	Pérez, Oliver	42414
3	SL	2022-04-24	76.9	2.35	5.38	Pérez, Oliver	42414
4	SI	2022-04-24	87.1	2.49	5.42	Pérez, Oliver	42414

5 rows × 21 columns

In [264]:
```
"""all_2022_stats_df.drop(columns=['spin_axis', 'effective_speed',
                                   'release_spin_rate', 'release_extension
"""
```

In [265]:
```
#all_2022_stats_df = all_2022_stats_df.dropna(axis=0)
```

In [266]:
```
#all_2022_stats_df.reset_index(inplace=True)
```

In [268]:
```
#all_2022_stats_df.drop('index', axis=1)
```

Out[268]:

	pitch_type	game_date	release_speed	release_pos_x	release_pos_z	player_name
0	SI	2022-04-24	87.8	2.74	5.20	Pérez, Oliver
1	SL	2022-04-24	75.8	2.12	5.60	Pérez, Oliver
2	SI	2022-04-24	87.3	2.45	5.41	Pérez, Oliver
3	SL	2022-04-24	76.9	2.35	5.38	Pérez, Oliver
4	SI	2022-04-24	87.1	2.49	5.42	Pérez, Oliver
...	...	...	...	...	...	...
329818	CH	2022-05-03	78.1	2.68	5.26	Yarbrough, Ryan
329819	FC	2022-05-03	83.1	3.05	5.03	Yarbrough, Ryan
329820	FC	2022-05-03	81.8	2.98	5.14	Yarbrough, Ryan
329821	CH	2022-05-03	77.8	2.60	5.39	Yarbrough, Ryan
329822	SI	2022-05-03	84.5	2.51	5.62	Yarbrough, Ryan

329823 rows × 17 columns

In [269]:

```
"""
# Group by 'game_date' and 'pitcher' to calculate the total pitches
total_pitches = all_2022_stats_df.groupby(['game_date', 'pitcher', 'pla

# Group by 'game_date', 'pitcher', and 'pitch_type' to calculate the su
total_pitches_by_type = all_2022_stats_df.groupby(['game_date', 'pitch

# Calculate averages of the specified metrics for each pitch type, grou
avg_metrics = all_2022_stats_df.groupby(['game_date', 'pitcher', 'play
    'release_speed': 'mean',
    'release_pos_x': 'mean',
    'release_pos_z': 'mean',
    'vx0': 'mean',
    'vy0': 'mean',
    'vz0': 'mean',
    'ax': 'mean',
    'ay': 'mean',
    'az': 'mean',
    'release_pos_y': 'mean',
}).reset_index()

grouped_2022_df = total_pitches.merge(total_pitches_by_type, on=['game_
grouped_2022_df = grouped_2022_df.merge(avg_metrics, on=['game_date',

grouped_2022_df.head()
"""
```

Out[269]:

	game_date	pitcher	player_name	total_pitches	pitch_type	count_by_pitch_type	release_
0	2022-04-07	424144	Pérez, Oliver	17	FF	1	89.9
1	2022-04-07	424144	Pérez, Oliver	17	SI	8	89.5
2	2022-04-07	424144	Pérez, Oliver	17	SL	8	77.2
3	2022-04-07	425794	Wainwright, Adam	81	CH	9	81.8
4	2022-04-07	425794	Wainwright, Adam	81	CS	2	69.1

In [270]:

```
"""
grouped_2022_df['game_date'] = pd.to_datetime(grouped_2022_df['game_da
grouped_2022_df['season'] = grouped_2022_df['game_date'].dt.year

# Step 1: Season Total Pitches
season_total_pitches = grouped_2022_df.groupby(['pitcher', 'player_nam

# Step 2: Season Total by Pitch Type
season_total_by_pitch_type = grouped_2022_df.groupby(['pitcher', 'play

# Weighted Averages Calculation Setup
weighted_avg_columns = ['release_speed', 'release_pos_x', 'release_pos_
for col in weighted_avg_columns:
    grouped_2022_df[f'{col}_product'] = grouped_2022_df[col] * grouped_

weighted_avg_aggregations = {f'{col}_product': 'sum' for col in weight
weighted_avg_aggregations['count_by_pitch_type'] = 'sum'

# Aggregate for weighted averages
weighted_avg_df = grouped_2022_df.groupby(['pitcher', 'player_name', '

# Calculate weighted averages
for col in weighted_avg_columns:
    weighted_avg_df[f'{col}_weighted_avg'] = weighted_avg_df[f'{col}_p

# Cleanup
weighted_avg_df.drop(columns=[f'{col}_product' for col in weighted_avg_

# Merge season totals and weighted averages
final_2022_df = pd.merge(season_total_pitches, season_total_by_pitch_ty
final_2022_df = pd.merge(final_2022_df, weighted_avg_df, on=['pitcher'

final_2022_df.head()
"""
```

Out[270]:

	pitcher	player_name	season	season_total_pitches	pitch_type	season_total_count_by_pit
0	424144	Pérez, Oliver	2022	179	FF	
1	424144	Pérez, Oliver	2022	179	SI	
2	424144	Pérez, Oliver	2022	179	SL	
3	425794	Wainwright, Adam	2022	17279	CH	
4	425794	Wainwright, Adam	2022	17279	CS	

In [271]:
```
"""
final_2022_df['player_name'] = final_2022_df['player_name'].str.lower(
final_2022_df.head()
"""
```

Out[271]:

	pitcher	player_name	season	season_total_pitches	pitch_type	season_total_count_by_pit
0	424144	pérez, oliver	2022	179	FF	
1	424144	pérez, oliver	2022	179	SI	
2	424144	pérez, oliver	2022	179	SL	
3	425794	wainwright, adam	2022	17279	CH	
4	425794	wainwright, adam	2022	17279	CS	

In [272]:
```
"""
def remove_accents(input_str):
    nfkd_form = unicodedata.normalize('NFKD', input_str)
    return "".join([c for c in nfkd_form if not unicodedata.combining(

def clean_name(name):
    name = name.lower()
    name = remove_accents(name)
    name = re.sub(r'[-.]', '', name)
    name = re.sub(r'\s+', ' ', name).strip()
    return name

final_2022_df['player_name'] = final_2022_df['player_name'].apply(clea
"""
```

In [273]:
```
"""
# Convert 'player_name' from "last name, first name" to "first name la
final_2022_df['Name'] = final_2022_df['player_name'].apply(lambda x: '
"""
```

In [274]:
```
"""
# Now, you can perform the merge using 'Name' and 'season' as the keys
better_2022_df = pd.merge(final_2022_df,
                          cleaning_filtered_df[['Name', 'season', 'Age
                          on=['Name', 'season'],
                          how='left')
"""
```

In [275]:
```
#better_2022_df.head()
```

Out[275]:

	pitcher	player_name	season	season_total_pitches	pitch_type	season_total_count_by_pit
0	424144	perez, oliver	2022	179	FF	
1	424144	perez, oliver	2022	179	SI	
2	424144	perez, oliver	2022	179	SL	
3	425794	wainwright, adam	2022	17279	CH	
4	425794	wainwright, adam	2022	17279	CS	

In [276]:
```
#better_2022_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 981 entries, 0 to 980
Data columns (total 19 columns):
 #   Column                            Non-Null Count  Dtype
---  ------                            --------------  -----
 0   pitcher                           981 non-null    int64
 1   player_name                       981 non-null    object
 2   season                            981 non-null    int32
 3   season_total_pitches              981 non-null    int64
 4   pitch_type                        981 non-null    object
 5   season_total_count_by_pitch_type  981 non-null    int64
 6   count_by_pitch_type               981 non-null    int64
 7   release_speed_weighted_avg        981 non-null    float64
 8   release_pos_x_weighted_avg        981 non-null    float64
 9   release_pos_z_weighted_avg        981 non-null    float64
 10  vx0_weighted_avg                  981 non-null    float64
 11  vy0_weighted_avg                  981 non-null    float64
 12  vz0_weighted_avg                  981 non-null    float64
 13  ax_weighted_avg                   981 non-null    float64
 14  ay_weighted_avg                   981 non-null    float64
 15  az_weighted_avg                   981 non-null    float64
 16  release_pos_y_weighted_avg        981 non-null    float64
 17  Name                              981 non-null    object
 18  Age                               882 non-null    float64
dtypes: float64(11), int32(1), int64(4), object(3)
memory usage: 141.9+ KB
```

In [277]:
```
#better_2022_df.to_csv('data/better_2022_df.csv')
```

2023

In [278]:
```
#all_2023_stats_df = pd.read_csv('~/Documents/Flatiron/Project_5_/data/
```

In [293]:
```
"""all_2023_stats_df.drop(columns=['batter', 'events', 'description',
                                   'des', 'game_type', 'stand', 'home_tea
                                   'away_team', 'type', 'hit_location', 'l
                                   'balls', 'strikes', 'pfx_x', 'spin_dir
                                   'pfx_z', 'plate_x', 'plate_z', 'on_3b'
                                   'on_2b', 'on_1b', 'outs_when_up', 'inn
                                   'inning_topbot', 'hc_x', 'hc_y', 'fiel
                                   'umpire', 'sv_id', 'hit_distance_sc',
                                    'sz_bot', 'launch_speed', 'launch_ang
                                   'pitcher.1', 'fielder_2.1', 'fielder_3
                                   'fielder_5', 'fielder_6', 'fielder_7',
                                   'fielder_9', 'estimated_ba_using_speed
                                   'estimated_woba_using_speedangle', 'bal
                                   'launch_speed_angle', 'woba_value', 'w
                                   'at_bat_number', 'pitch_number', 'home_
                                   'bat_score', 'fld_score', 'post_home_s
                                   'post_fld_score', 'post_away_score', '
                                   'of_fielding_alignment', 'delta_home_w
                                   'delta_run_exp', 'spin_rate_deprecated
                                  'break_length_deprecated', 'tfs_deprecat
all_2023_stats_df.head()
"""
```

Out[293]:

	pitch_type	game_date	release_speed	release_pos_x	release_pos_z	player_name	pitche
0	FA	2023-04-14	52.1	-1.99	6.71	Pérez, Carlos	54220
1	FA	2023-04-14	69.6	-2.30	6.58	Pérez, Carlos	54220
2	FA	2023-04-14	65.1	-2.25	6.52	Pérez, Carlos	54220
3	FA	2023-04-14	64.2	-2.37	6.50	Pérez, Carlos	54220
4	FA	2023-04-14	72.0	-2.45	6.49	Pérez, Carlos	54220

5 rows × 21 columns

In [294]:
```
"""
all_2023_stats_df.drop(columns=['spin_axis', 'effective_speed',
                                'release_spin_rate', 'release_extension
"""
```

In [295]:
```
#all_2023_stats_df = all_2023_stats_df.dropna(axis=0)
```

In [296]:
```
#all_2023_stats_df.reset_index(inplace=True)
```

In [297]:
```
#all_2023_stats_df.drop('index', axis=1)
```

Out[297]:

	pitch_type	game_date	release_speed	release_pos_x	release_pos_z	player_name
0	FA	2023-04-14	52.1	-1.99	6.71	Pérez, Carlos
1	FA	2023-04-14	69.6	-2.30	6.58	Pérez, Carlos
2	FA	2023-04-14	65.1	-2.25	6.52	Pérez, Carlos
3	FA	2023-04-14	64.2	-2.37	6.50	Pérez, Carlos
4	FA	2023-04-14	72.0	-2.45	6.49	Pérez, Carlos
...	...	...	...	...	...	...
290106	CU	2023-04-01	70.1	3.02	5.35	Yarbrough, Ryan
290107	CU	2023-04-01	71.3	2.90	5.33	Yarbrough, Ryan
290108	CH	2023-04-01	79.6	2.71	5.48	Yarbrough, Ryan
290109	SI	2023-04-01	87.8	2.84	5.59	Yarbrough, Ryan
290110	SI	2023-04-01	86.6	2.88	5.61	Yarbrough, Ryan

290111 rows × 17 columns

In [298]:

```
"""# Group by 'game_date' and 'pitcher' to calculate the total pitches
total_pitches = all_2023_stats_df.groupby(['game_date', 'pitcher', 'pl

# Group by 'game_date', 'pitcher', and 'pitch_type' to calculate the s
total_pitches_by_type = all_2023_stats_df.groupby(['game_date', 'pitch

# Calculate averages of the specified metrics for each pitch type, gro
avg_metrics = all_2023_stats_df.groupby(['game_date', 'pitcher', 'play
    'release_speed': 'mean',
    'release_pos_x': 'mean',
    'release_pos_z': 'mean',
    'vx0': 'mean',
    'vy0': 'mean',
    'vz0': 'mean',
    'ax': 'mean',
    'ay': 'mean',
    'az': 'mean',
    'release_pos_y': 'mean',
}).reset_index()

grouped_2023_df = total_pitches.merge(total_pitches_by_type, on=['game_
grouped_2023_df = grouped_2023_df.merge(avg_metrics, on=['game_date',

grouped_2023_df.head()
"""
```

Out[298]:

	game_date	pitcher	player_name	total_pitches	pitch_type	count_by_pitch_type	release_
0	2023-03-30	425844	Greinke, Zack	80	CH	11	87.6
1	2023-03-30	425844	Greinke, Zack	80	CU	22	73.8
2	2023-03-30	425844	Greinke, Zack	80	FC	9	86.
3	2023-03-30	425844	Greinke, Zack	80	FF	6	90.3
4	2023-03-30	425844	Greinke, Zack	80	SI	13	90.7

In [299]:

```
"""
grouped_2023_df['game_date'] = pd.to_datetime(grouped_2023_df['game_da
grouped_2023_df['season'] = grouped_2023_df['game_date'].dt.year

# Step 1: Season Total Pitches
season_total_pitches = grouped_2023_df.groupby(['pitcher', 'player_nam

# Step 2: Season Total by Pitch Type
season_total_by_pitch_type = grouped_2023_df.groupby(['pitcher', 'play

# Weighted Averages Calculation Setup
weighted_avg_columns = ['release_speed', 'release_pos_x', 'release_pos_
for col in weighted_avg_columns:
    grouped_2023_df[f'{col}_product'] = grouped_2023_df[col] * grouped_

weighted_avg_aggregations = {f'{col}_product': 'sum' for col in weight
weighted_avg_aggregations['count_by_pitch_type'] = 'sum'

# Aggregate for weighted averages
weighted_avg_df = grouped_2023_df.groupby(['pitcher', 'player_name', '

# Calculate weighted averages
for col in weighted_avg_columns:
    weighted_avg_df[f'{col}_weighted_avg'] = weighted_avg_df[f'{col}_p

# Cleanup
weighted_avg_df.drop(columns=[f'{col}_product' for col in weighted_avg_

# Merge season totals and weighted averages
final_2023_df = pd.merge(season_total_pitches, season_total_by_pitch_ty
final_2023_df = pd.merge(final_2023_df, weighted_avg_df, on=['pitcher'

final_2023_df.head()
"""
```

Out[299]:

	pitcher	player_name	season	season_total_pitches	pitch_type	season_total_count_by_pit
0	425794	Wainwright, Adam	2023	9498	CH	
1	425794	Wainwright, Adam	2023	9498	CS	
2	425794	Wainwright, Adam	2023	9498	CU	
3	425794	Wainwright, Adam	2023	9498	FC	
4	425794	Wainwright, Adam	2023	9498	FF	

In [300]:
```
"""
final_2023_df['player_name'] = final_2023_df['player_name'].str.lower(
final_2023_df.head()
"""
```

Out[300]:

	pitcher	player_name	season	season_total_pitches	pitch_type	season_total_count_by_pit
0	425794	wainwright, adam	2023	9498	CH	
1	425794	wainwright, adam	2023	9498	CS	
2	425794	wainwright, adam	2023	9498	CU	
3	425794	wainwright, adam	2023	9498	FC	
4	425794	wainwright, adam	2023	9498	FF	

In [301]:
```
"""
def remove_accents(input_str):
    nfkd_form = unicodedata.normalize('NFKD', input_str)
    return "".join([c for c in nfkd_form if not unicodedata.combining(

def clean_name(name):
    name = name.lower()
    name = remove_accents(name)
    name = re.sub(r'[-.]', '', name)
    name = re.sub(r'\s+', ' ', name).strip()
    return name

final_2023_df['player_name'] = final_2023_df['player_name'].apply(clea
"""
```

In [302]:
```
"""
# Convert 'player_name' from "last name, first name" to "first name la
final_2023_df['Name'] = final_2023_df['player_name'].apply(lambda x: '
"""
```

In [303]:
```
"""
# Now, you can perform the merge using 'Name' and 'season' as the keys
better_2023_df = pd.merge(final_2023_df,
                          cleaning_filtered_df[['Name', 'season', 'Age
                          on=['Name', 'season'],
                          how='left')
"""
```

In [304]:
```
#better_2023_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 877 entries, 0 to 876
Data columns (total 19 columns):
 #   Column                           Non-Null Count  Dtype
---  ------                           --------------  -----
 0   pitcher                          877 non-null    int64
 1   player_name                      877 non-null    object
 2   season                           877 non-null    int32
 3   season_total_pitches             877 non-null    int64
 4   pitch_type                       877 non-null    object
 5   season_total_count_by_pitch_type 877 non-null    int64
 6   count_by_pitch_type              877 non-null    int64
 7   release_speed_weighted_avg       877 non-null    float64
 8   release_pos_x_weighted_avg       877 non-null    float64
 9   release_pos_z_weighted_avg       877 non-null    float64
 10  vx0_weighted_avg                 877 non-null    float64
 11  vy0_weighted_avg                 877 non-null    float64
 12  vz0_weighted_avg                 877 non-null    float64
 13  ax_weighted_avg                  877 non-null    float64
 14  ay_weighted_avg                  877 non-null    float64
 15  az_weighted_avg                  877 non-null    float64
 16  release_pos_y_weighted_avg       877 non-null    float64
 17  Name                             877 non-null    object
 18  Age                              815 non-null    float64
dtypes: float64(11), int32(1), int64(4), object(3)
memory usage: 126.9+ KB
```

In [305]:
```
#better_2023_df.head()
```

Out[305]:

	pitcher	player_name	season	season_total_pitches	pitch_type	season_total_count_by_pit
0	425794	wainwright, adam	2023	9498	CH	
1	425794	wainwright, adam	2023	9498	CS	
2	425794	wainwright, adam	2023	9498	CU	
3	425794	wainwright, adam	2023	9498	FC	
4	425794	wainwright, adam	2023	9498	FF	

In [306]:
```
#better_2023_df.to_csv('data/better_2023_df.csv')
```

END.