

```
In [18]: 1 import pybaseball as pyb
2 from pybaseball import statcast, pitching_stats, playerid_lookup, stat
3 import numpy as np
4 import math
5 import pandas as pd
6 import glob
7 import os
8 import re
9 import unicodedata
10 from datetime import datetime
11 from itertools import groupby
12 from operator import itemgetter
13 from sklearn.preprocessing import OneHotEncoder
```

Load in all 'better' DF and concat.

```
In [19]: 1 better_2023_df = pd.read_csv('better_2023_df.csv', index_col=0)
2 better_2022_df = pd.read_csv('better_2022_df.csv', index_col=0)
3 better_2021_df = pd.read_csv('better_2021_df.csv', index_col=0)
4 better_2020_df = pd.read_csv('better_2020_df.csv', index_col=0)
5 better_2019_df = pd.read_csv('better_2019_df.csv', index_col=0)
6 better_2018_df = pd.read_csv('better_2018_df.csv', index_col=0)
7 better_2017_df = pd.read_csv('better_2017_df.csv', index_col=0)
8 better_2016_df = pd.read_csv('better_2016_df.csv', index_col=0)
9 better_2015_df = pd.read_csv('better_2015_df.csv', index_col=0)
10 better_2014_df = pd.read_csv('better_2014_df.csv', index_col=0)
11 better_2013_df = pd.read_csv('better_2013_df.csv', index_col=0)
12 better_2012_df = pd.read_csv('better_2012_df.csv', index_col=0)
13 better_2011_df = pd.read_csv('better_2011_df.csv', index_col=0)
14 better_2010_df = pd.read_csv('better_2010_df.csv', index_col=0)
15 better_2009_df = pd.read_csv('better_2009_df.csv', index_col=0)
16 better_2008_df = pd.read_csv('better_2008_df.csv', index_col=0)
```

In [20]: 1 better\_2023\_df.info()

```
<class 'pandas.core.frame.DataFrame'>
```

```
Index: 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	int64
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), int64(5), object(3)
```

```
memory usage: 137.0+ KB
```

In [21]: 1 better\_2015\_df.info()

<class 'pandas.core.frame.DataFrame'>

Index: 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	int64
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), int64(5), object(3)

memory usage: 230.5+ KB

In [22]: 1 better\_2010\_df.info()

```
<class 'pandas.core.frame.DataFrame'>
Index: 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   int64
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), int64(5), object(3)
memory usage: 248.9+ KB
```

In [28]: 1 dataframes = [  
2 better\_2023\_df, better\_2022\_df, better\_2021\_df, better\_2020\_df,  
3 better\_2019\_df, better\_2018\_df, better\_2017\_df, better\_2016\_df,  
4 better\_2015\_df, better\_2014\_df, better\_2013\_df, better\_2012\_df,  
5 better\_2011\_df, better\_2010\_df, better\_2009\_df, better\_2008\_df  
6 ]  
7  
8 total\_years\_df = pd.concat(dataframes, ignore\_index=True)

In [29]: 1 total\_years\_df.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 21398 entries, 0 to 21397
Data columns (total 19 columns):
#   Column                                          Non-Null Count  Dtype
---  -
0   pitcher                                         21398 non-null   int64
1   player_name                                    21398 non-null   object
2   season                                          21398 non-null   int64
3   season_total_pitches                          21398 non-null   int64
4   pitch_type                                    21398 non-null   object
5   season_total_count_by_pitch_type             21398 non-null   int64
6   count_by_pitch_type                          21398 non-null   int64
7   release_speed_weighted_avg                   21398 non-null   float64
8   release_pos_x_weighted_avg                   21398 non-null   float64
9   release_pos_z_weighted_avg                   21398 non-null   float64
10  vx0_weighted_avg                             21398 non-null   float64
11  vy0_weighted_avg                             21398 non-null   float64
12  vz0_weighted_avg                             21398 non-null   float64
13  ax_weighted_avg                              21398 non-null   float64
14  ay_weighted_avg                              21398 non-null   float64
15  az_weighted_avg                              21398 non-null   float64
16  release_pos_y_weighted_avg                   21398 non-null   float64
17  Name                                           21398 non-null   object
18  Age                                           18652 non-null   float64
dtypes: float64(11), int64(5), object(3)
memory usage: 3.1+ MB
```

In [30]: 1 total\_years\_df.head()

Out[30]:

	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	

Drop ax, ay, az, player\_name, count\_by\_pitch\_type.

Reorganize DF so Name and Age are at the front

Move release\_pos\_y\_weighted\_avg near like columns.

Make pitch\_type into features.

```
In [34]: 1 total_years_df.drop(columns=['player_name', 'count_by_pitch_type', 'ax',
2                                     'ay_weighted_avg', 'az_weighted_avg'], in
```

```
In [35]: 1 total_years_df.head()
```

```
Out[35]:
```

	pitcher	season	season_total_pitches	pitch_type	season_total_count_by_pitch_type	relea
0	425794	2023	9498	CH		91
1	425794	2023	9498	CS		3
2	425794	2023	9498	CU		545
3	425794	2023	9498	FC		403
4	425794	2023	9498	FF		176

```
In [39]: 1 total_years_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 21398 entries, 0 to 21397
Data columns (total 14 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   Name                                  21398 non-null  object
1   Age                                   18652 non-null  float64
2   pitcher                              21398 non-null  int64
3   season                               21398 non-null  int64
4   season_total_pitches                 21398 non-null  int64
5   pitch_type                           21398 non-null  object
6   season_total_count_by_pitch_type     21398 non-null  int64
7   release_speed_weighted_avg           21398 non-null  float64
8   release_pos_x_weighted_avg           21398 non-null  float64
9   release_pos_y_weighted_avg           21398 non-null  float64
10  release_pos_z_weighted_avg           21398 non-null  float64
11  vx0_weighted_avg                     21398 non-null  float64
12  vy0_weighted_avg                     21398 non-null  float64
13  vz0_weighted_avg                     21398 non-null  float64
dtypes: float64(8), int64(4), object(2)
memory usage: 2.3+ MB
```

```
In [37]: 1 new_order = [  
2     'Name', 'Age', 'pitcher', 'season', 'season_total_pitches', 'pitch  
3     'season_total_count_by_pitch_type', 'release_speed_weighted_avg',  
4     'release_pos_y_weighted_avg', 'release_pos_z_weighted_avg', 'vx0_w  
5     'vy0_weighted_avg', 'vz0_weighted_avg'  
6 ]  
7  
8 total_years_df = total_years_df[new_order]
```

```
In [38]: 1 total_years_df.head()
```

Out[38]:

	Name	Age	pitcher	season	season_total_pitches	pitch_type	season_total_count_by
--	------	-----	---------	--------	----------------------	------------	-----------------------

0	adam wainwright	41.0	425794	2023	9498	CH	
1	adam wainwright	41.0	425794	2023	9498	CS	
2	adam wainwright	41.0	425794	2023	9498	CU	
3	adam wainwright	41.0	425794	2023	9498	FC	
4	adam wainwright	41.0	425794	2023	9498	FF	

```
In [841]: 1 # Remove all apostrophies from names  
2 total_years_df['Name'] = total_years_df['Name'].str.replace("'", "")  
3 print(total_years_df[total_years_df['Name'].str.contains('sean osulliv
```

...

```
In [1023]: 1 # Filter the DataFrame to show only rows where 'Age' is NaN
           2 nan_age_rows = total_years_df[total_years_df['Age'].isna()]
           3 nan_age_rows
```

Out[1023]:

	Name	Age	pitcher	season	season_total_pitches	pitch_type	season_total_count_b
--	------	-----	---------	--------	----------------------	------------	----------------------

155	carlos perez	NaN	542208	2023	67	EP	
156	carlos perez	NaN	542208	2023	67	FA	
20169	damian moss	NaN	150305	2008	185	CH	
20170	damian moss	NaN	150305	2008	185	CU	
20171	damian moss	NaN	150305	2008	185	FF	
20172	damian moss	NaN	150305	2008	185	SI	
20173	damian moss	NaN	150305	2008	185	SL	

This part took a lot of manual entry of searching for names and ages, cross-referencing with [baseball-reference.com](http://baseball-reference.com)

Only two pitchers were not found, so they were dropped from the DF.

```
In [1021]: 1 age_to_fill = 32 # Starting age
           2 start_season = 2008 # Starting season
           3
           4 # Loop from start_season down to 2008, decrementing the age and season
           5 for season in range(start_season, 2007, -1): # End at 2008
           6     total_years_df.loc[(total_years_df['Name'] == 'scott elarton') & (
           7         age_to_fill -= 1 # Decrement the age for the next iteration
           8
           9 # Verify the changes for a range of seasons to see if the loop worked
          10 print(total_years_df[(total_years_df['Name'] == 'scott elarton') & (to
          11
```

	Name	season	Age
20086	scott elarton	2008	32.0
20087	scott elarton	2008	32.0
20088	scott elarton	2008	32.0
20089	scott elarton	2008	32.0
20090	scott elarton	2008	32.0
20091	scott elarton	2008	32.0

```
In [1024]: 1 complete_df = total_years_df[total_years_df['Name'] != 'carlos perez']
```



```
In [1027]: 1 complete_df = complete_df[complete_df['Name'] != 'damian moss']
```

```
In [1028]: 1 complete_df.info()
```

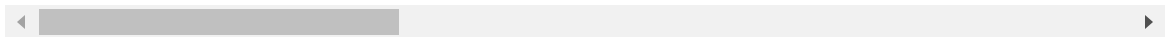
```
<class 'pandas.core.frame.DataFrame'>
Index: 21391 entries, 0 to 21397
Data columns (total 14 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   Name                                21391 non-null  object
1   Age                                21391 non-null  float64
2   pitcher                            21391 non-null  int64
3   season                             21391 non-null  int64
4   season_total_pitches               21391 non-null  int64
5   pitch_type                         21391 non-null  object
6   season_total_count_by_pitch_type  21391 non-null  int64
7   release_speed_weighted_avg        21391 non-null  float64
8   release_pos_x_weighted_avg        21391 non-null  float64
9   release_pos_y_weighted_avg        21391 non-null  float64
10  release_pos_z_weighted_avg        21391 non-null  float64
11  vx0_weighted_avg                  21391 non-null  float64
12  vy0_weighted_avg                  21391 non-null  float64
13  vz0_weighted_avg                  21391 non-null  float64
dtypes: float64(8), int64(4), object(2)
memory usage: 2.4+ MB
```

In [1032]: 1 complete\_df

Out[1032]:

	Name	Age	pitcher	season	season_total_pitches	pitch_type	season_total_cour
0	adam wainwright	41.0	425794	2023	9498	CH	
1	adam wainwright	41.0	425794	2023	9498	CS	
2	adam wainwright	41.0	425794	2023	9498	CU	
3	adam wainwright	41.0	425794	2023	9498	FC	
4	adam wainwright	41.0	425794	2023	9498	FF	
...	...	...	...	...	...	...	...
21393	jeff samardzija	23.0	502188	2008	1820	FS	
21394	jeff samardzija	23.0	502188	2008	1820	IN	
21395	jeff samardzija	23.0	502188	2008	1820	PO	
21396	jeff samardzija	23.0	502188	2008	1820	SI	
21397	jeff samardzija	23.0	502188	2008	1820	SL	

21391 rows × 14 columns



In [1029]: 1 complete\_df.to\_csv('complete\_df.csv')

From here, need to make 'pitch\_type' into features.

Also need to merge w/ pitcher\_key\_df.csv to add back in:

'Throws', 'Surgery', 'TJ Surgery Year'

In [3]: 1 complete\_df = pd.read\_csv('complete\_df.csv', index\_col=0)

In [4]: 1 pitcher\_data\_df = pd.read\_csv('pitcher\_key\_df.csv', index\_col=0)

In [5]: 1 pitcher\_data\_df.info()

```
<class 'pandas.core.frame.DataFrame'>
Index: 1242 entries, 240 to 15350
Data columns (total 21 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Name                   1242 non-null   object
1   Age                    1242 non-null   int64
2   Year                   1242 non-null   int64
3   Throws                 1242 non-null   int64
4   IP                     1242 non-null   float64
5   G                      1242 non-null   int64
6   GS                     1242 non-null   int64
7   CG                     1242 non-null   int64
8   SHO                    1242 non-null   int64
9   SDR                    1242 non-null   int64
10  Career Start           1242 non-null   int64
11  Career End             1242 non-null   int64
12  Inactive Years         1242 non-null   object
13  Surgery                1242 non-null   float64
14  TJ Surgery Date        268 non-null    object
15  Surgeon(s)             197 non-null    object
16  Country                 268 non-null    object
17  Level                   268 non-null    object
18  Total_IP               1242 non-null   float64
19  TJ Surgery Year         268 non-null    float64
20  key_mlbam              1242 non-null   int64
dtypes: float64(4), int64(11), object(6)
memory usage: 213.5+ KB
```

In [6]: 1 complete\_100\_df = pd.merge(complete\_df,  
2 pitcher\_data\_df[['key\_mlbam', 'Throws', 'Su  
3 left\_on='pitcher',  
4 right\_on='key\_mlbam',  
5 how='left'])

In [7]: 1 complete\_100\_df.columns

```
Out[7]: Index(['Name', 'Age', 'pitcher', 'season', 'season_total_pitches',
'pitch_type', 'season_total_count_by_pitch_type',
'release_speed_weighted_avg', 'release_pos_x_weighted_avg',
'release_pos_y_weighted_avg', 'release_pos_z_weighted_avg',
'vx0_weighted_avg', 'vy0_weighted_avg', 'vz0_weighted_avg', 'key_m
lbam',
'Throws', 'Surgery', 'TJ Surgery Year'],
dtype='object')
```

In [8]: 1 complete\_100\_df.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 21391 entries, 0 to 21390
Data columns (total 18 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   Name                                21391 non-null  object
1   Age                                21391 non-null  float64
2   pitcher                            21391 non-null  int64
3   season                             21391 non-null  int64
4   season_total_pitches               21391 non-null  int64
5   pitch_type                         21391 non-null  object
6   season_total_count_by_pitch_type  21391 non-null  int64
7   release_speed_weighted_avg         21391 non-null  float64
8   release_pos_x_weighted_avg         21391 non-null  float64
9   release_pos_y_weighted_avg         21391 non-null  float64
10  release_pos_z_weighted_avg         21391 non-null  float64
11  vx0_weighted_avg                   21391 non-null  float64
12  vy0_weighted_avg                   21391 non-null  float64
13  vz0_weighted_avg                   21391 non-null  float64
14  key_mlbam                          21391 non-null  int64
15  Throws                             21391 non-null  int64
16  Surgery                             21391 non-null  float64
17  TJ Surgery Year                     7579 non-null   float64
dtypes: float64(10), int64(6), object(2)
memory usage: 2.9+ MB
```

In [9]: 1 complete\_100\_df.head()

Out[9]:

	Name	Age	pitcher	season	season_total_pitches	pitch_type	season_total_count_by.
0	adam wainwright	41.0	425794	2023	9498	CH	
1	adam wainwright	41.0	425794	2023	9498	CS	
2	adam wainwright	41.0	425794	2023	9498	CU	
3	adam wainwright	41.0	425794	2023	9498	FC	
4	adam wainwright	41.0	425794	2023	9498	FF	

In [10]: 1 complete\_100\_df.drop(columns=['key\_mlbam'], inplace=True)

In [11]: 1 complete\_100\_df.info()

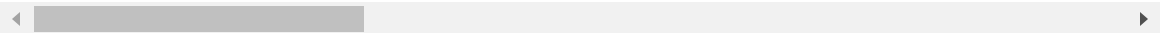
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 21391 entries, 0 to 21390
Data columns (total 17 columns):
#   Column                                     Non-Null Count  Dtype
---  -
0   Name                                       21391 non-null  object
1   Age                                       21391 non-null  float64
2   pitcher                                   21391 non-null  int64
3   season                                   21391 non-null  int64
4   season_total_pitches                     21391 non-null  int64
5   pitch_type                               21391 non-null  object
6   season_total_count_by_pitch_type         21391 non-null  int64
7   release_speed_weighted_avg               21391 non-null  float64
8   release_pos_x_weighted_avg               21391 non-null  float64
9   release_pos_y_weighted_avg               21391 non-null  float64
10  release_pos_z_weighted_avg               21391 non-null  float64
11  vx0_weighted_avg                         21391 non-null  float64
12  vy0_weighted_avg                         21391 non-null  float64
13  vz0_weighted_avg                         21391 non-null  float64
14  Throws                                   21391 non-null  int64
15  Surgery                                   21391 non-null  float64
16  TJ Surgery Year                           7579 non-null   float64
dtypes: float64(10), int64(5), object(2)
memory usage: 2.8+ MB
```

In [12]: 1 complete\_100\_df

Out[12]:

	Name	Age	pitcher	season	season_total_pitches	pitch_type	season_total_cour
0	adam wainwright	41.0	425794	2023	9498	CH	
1	adam wainwright	41.0	425794	2023	9498	CS	
2	adam wainwright	41.0	425794	2023	9498	CU	
3	adam wainwright	41.0	425794	2023	9498	FC	
4	adam wainwright	41.0	425794	2023	9498	FF	
...	...	...	...	...	...	...	
21386	jeff samardzija	23.0	502188	2008	1820	FS	
21387	jeff samardzija	23.0	502188	2008	1820	IN	
21388	jeff samardzija	23.0	502188	2008	1820	PO	
21389	jeff samardzija	23.0	502188	2008	1820	SI	
21390	jeff samardzija	23.0	502188	2008	1820	SL	

21391 rows × 17 columns



In [1061]: 1 pd.set\_option('display.max\_rows', None)  
2 pd.set\_option('display.max\_columns', None)

Something went wrong with 'season\_total\_pitches', must drop.

'season\_total\_count\_by\_pitch\_type' is correct. Thats what matters.

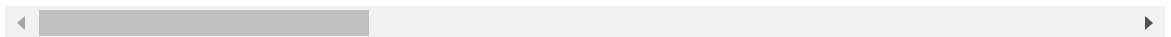
In [13]: 1 adam\_df = complete\_100\_df[complete\_100\_df['Name'] == 'adam wainwright']

In [14]: 1 adam\_df

Out[14]:

	Name	Age	pitcher	season	season_total_pitches	pitch_type	season_total_coun
0	adam wainwright	41.0	425794	2023	9498	CH	
1	adam wainwright	41.0	425794	2023	9498	CS	
2	adam wainwright	41.0	425794	2023	9498	CU	
3	adam wainwright	41.0	425794	2023	9498	FC	
4	adam wainwright	41.0	425794	2023	9498	FF	
...	...	...	...	...	...	...	
20627	adam wainwright	26.0	425794	2008	8850	FC	
20628	adam wainwright	26.0	425794	2008	8850	FF	
20629	adam wainwright	26.0	425794	2008	8850	IN	
20630	adam wainwright	26.0	425794	2008	8850	PO	
20631	adam wainwright	26.0	425794	2008	8850	SI	

95 rows × 17 columns



In [15]: 1 complete\_100\_df['Surgery'].value\_counts()

Out[15]: Surgery  
 0.0 13812  
 1.0 6629  
 2.0 930  
 3.0 20  
 Name: count, dtype: int64

In [16]: 1 complete\_100\_df.loc[complete\_100\_df['Surgery'] != 0.0, 'Surgery'] = 1.0

In [17]: 1 complete\_100\_df.to\_csv('complete\_100\_df.csv')