In [18]:

```
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
```

Load in all 'better' DF and concat.

In [19]:

```
better_2023_df = pd.read_csv('~/Documents/Flatiron/Project_5_/data/bet
better_2022_df = pd.read_csv('~/Documents/Flatiron/Project_5_/data/bet
better_2021_df = pd.read_csv('~/Documents/Flatiron/Project_5_/data/bet
better_2020_df = pd.read_csv('~/Documents/Flatiron/Project_5_/data/bet
better_2019_df = pd.read_csv('~/Documents/Flatiron/Project_5_/data/bet
better_2018_df = pd.read_csv('~/Documents/Flatiron/Project_5_/data/bet
better_2017_df = pd.read_csv('~/Documents/Flatiron/Project_5_/data/bet
better_2016_df = pd.read_csv('~/Documents/Flatiron/Project_5_/data/bet
better_2015_df = pd.read_csv('~/Documents/Flatiron/Project_5_/data/bet
better_2014_df = pd.read_csv('~/Documents/Flatiron/Project_5_/data/bet
better_2013_df = pd.read_csv('~/Documents/Flatiron/Project_5_/data/bet
better_2012_df = pd.read_csv('~/Documents/Flatiron/Project_5_/data/bet
better_2011_df = pd.read_csv('~/Documents/Flatiron/Project_5_/data/bet
better_2010_df = pd.read_csv('~/Documents/Flatiron/Project_5_/data/bet
better_2009_df = pd.read_csv('~/Documents/Flatiron/Project_5_/data/bet
better_2008_df = pd.read_csv('~/Documents/Flatiron/Project_5_/data/bet
```

In [20]:

```
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]:

```
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]:

```python
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]:

```python
dataframes = [
    better_2023_df, better_2022_df, better_2021_df, better_2020_df,
    better_2019_df, better_2018_df, better_2017_df, better_2016_df,
    better_2015_df, better_2014_df, better_2013_df, better_2012_df,
    better_2011_df, better_2010_df, better_2009_df, better_2008_df
]

total_years_df = pd.concat(dataframes, ignore_index=True)
```

In [29]:

```
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]:

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

In [35]:
```
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]:
```
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]:

```
new_order = [
    'Name', 'Age', 'pitcher', 'season', 'season_total_pitches', 'pitch_
    'season_total_count_by_pitch_type', 'release_speed_weighted_avg',
    'release_pos_y_weighted_avg', 'release_pos_z_weighted_avg', 'vx0_w
    'vy0_weighted_avg', 'vz0_weighted_avg'
]

total_years_df = total_years_df[new_order]
```

In [38]:

```
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]:

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

```
                Name
7870   sean osullivan
7871   sean osullivan
7872   sean osullivan
7873   sean osullivan
7874   sean osullivan
9425   sean osullivan
9426   sean osullivan
9427   sean osullivan
9428   sean osullivan
9429   sean osullivan
9430   sean osullivan
10975  sean osullivan
10976  sean osullivan
10977  sean osullivan
10978  sean osullivan
10979  sean osullivan
12613  sean osullivan
12614  sean osullivan
12615  sean osullivan
12616  sean osullivan
12617  sean osullivan
12618  sean osullivan
12619  sean osullivan
12620  sean osullivan
14319  sean osullivan
14320  sean osullivan
14321  sean osullivan
14322  sean osullivan
14323  sean osullivan
16030  sean osullivan
16031  sean osullivan
16032  sean osullivan
16033  sean osullivan
16034  sean osullivan
16035  sean osullivan
16036  sean osullivan
17794  sean osullivan
17795  sean osullivan
17796  sean osullivan
17797  sean osullivan
17798  sean osullivan
17799  sean osullivan
17800  sean osullivan
19522  sean osullivan
19523  sean osullivan
19524  sean osullivan
19525  sean osullivan
19526  sean osullivan
19527  sean osullivan
19528  sean osullivan
```

In [1023]:

```python
# Filter the DataFrame to show only rows where 'Age' is NaN
nan_age_rows = total_years_df[total_years_df['Age'].isna()]
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

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

In [1021]:

```python
age_to_fill = 32 # Starting age
start_season = 2008  # Starting season

# Loop from start_season down to 2008, decrementing the age and season
for season in range(start_season, 2007, -1):  # End at 2008
    total_years_df.loc[(total_years_df['Name'] == 'scott elarton') & (
    age_to_fill -= 1  # Decrement the age for the next iteration

# Verify the changes for a range of seasons to see if the loop worked 
print(total_years_df[(total_years_df['Name'] == 'scott elarton') & (to

```

```
                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]:

```python
complete_df = total_years_df[total_years_df['Name'] != 'carlos perez']
```

In [1027]:

```
complete_df = complete_df[complete_df['Name'] != 'damian moss']
```

In [1028]:

```
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]:

```
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]:

```
complete_df.to_csv('data/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 [4]:

```
pitcher_data_df = pd.read_csv('~/Documents/Flatiron/Project_5_/data/pi
```

In [5]:

```
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]:

```
complete_100_df = pd.merge(complete_df,
                           pitcher_data_df[['key_mlbam', 'Throws', 'Su
                           left_on='pitcher',
                           right_on='key_mlbam',
                           how='left')
```

In [7]:

```
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]:

```
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]:

```
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]:

```
complete_100_df.drop(columns=['key_mlbam'], inplace=True)
```

In [11]:

```
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]:

```
complete_100_df
```

Out[12]:

	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	
...	...	...	...	...	...	...	
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]:

```
pd.set_option('display.max_rows', None)
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]:

```
adam_df = complete_100_df[complete_100_df['Name'] == 'adam wainwright'
```

In [14]:

```
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]:

```
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]:

```
complete_100_df.loc[complete_100_df['Surgery'] != 0.0, 'Surgery'] = 1.
```

In [17]:

```
complete_100_df.to_csv('data/complete_100_df.csv')
```