Database Data Analysis

Nathan Shaver

Preface

Sabermetrics is the empirical analysis of baseball, especially baseball statistics that measure in-game activity. These activities measure pitching, batting, offensive and defensive metrics. Analysis of these statistics has resulted in more advanced statistical measurements that coaches and front offices use to analyze players in a unique pattern that simpler statistics wouldn't offer. The development of Sabermetrics statistics wouldn't be possible without the development of programming and data analysis.

For example:

Classic Pitching Statistics

- ERA (Earned Run Avg)
- IP (Innings Pitched)
- BB (Base on Balls)
- K (Strikeouts)
- H (Hits)

SABR Pitching Statistics

- WHIP (Walks plus hits per Innings Pitched)
- BABIP (Batting avg on balls in play)
- FiP (Fielding Independent Pitching)
- SIERA (Skill-interactive Earned Run Avg)
- BQR (Bequeathed Runners Scored)

Premise of the project

pybaseball is a Python library for analyzing baseball data. The library pulls data from Baseball Reference, Baseball Savant, and FanGraphs, official websites that store tremendous amounts of information. The software gets statcast data, pitching statistics, hitting statistics, division standings/team records, awards data, and other information. Data is accessible at the individual pitching/batting level as well as aggregated at the season and time period level.

References

The package was originally developed by James LeDoux and is maintained by Moshe Schorr. This package was inspired by Bill Petti's R package baseballr, which at the time of the package's development had no Python equivalent. There are currently 27 contributors and is free of charge for anyone to use.

Import statcast

Statcast data include pitch-level features such as Perceived Velocity (PV), Spin Rate (SR), Exit Velocity (EV), pitch X, Y, and Z coordinates, and more. The function statcast(start_dt, end_dt) pulls this data from baseballsavant.com.

```
>>> from pybaseball import statcast
>>> data = statcast(start_dt='2017-06-24', end_dt='2017-06-27')
>>> data.head(2)

   index pitch_type  game_date  release_speed  release_pos_x  release_pos_z
0    314         CU 2017-06-27           79.7        -1.3441         5.4075
1    332         FF 2017-06-27           98.1        -1.3547         5.4196

  player_name    batter   pitcher     events     ...      release_pos_y
0   Matt Bush  608070.0  456713.0  field_out     ...            54.8585
1   Matt Bush  429665.0  456713.0  field_out     ...            54.3470

   estimated_ba_using_speedangle  estimated_woba_using_speedangle  woba_value
0                          0.100                            0.137         0.0
1                          0.269                            0.258         0.0

   woba_denom babip_value iso_value launch_speed_angle at_bat_number pitch_number
0         1.0         0.0       0.0                3.0          64.0          1.0
1         1.0         0.0       0.0                3.0          63.0          3.0
[2 rows x 79 columns]
```

Import pitching_stats

Import includes pitching stats for players across multiple seasons, single seasons, or during a specified time period. Each season has roughly 700,000 pitches and updates every year.

```
>>> from pybaseball import pitching_stats
>>> data = pitching_stats(2012, 2016)
>>> data.head()
     Season            Name     Team   Age     W    L   ERA  WAR     G    GS
336  2015.0  Clayton Kershaw  Dodgers  27.0  16.0  7.0  2.13  8.6  33.0  33.0
236  2014.0  Clayton Kershaw  Dodgers  26.0  21.0  3.0  1.77  7.6  27.0  27.0
472  2014.0     Corey Kluber  Indians  28.0  18.0  9.0  2.44  7.4  34.0  34.0
235  2015.0     Jake Arrieta     Cubs  29.0  22.0  6.0  1.77  7.3  33.0  33.0
256  2013.0  Clayton Kershaw  Dodgers  25.0  16.0  9.0  1.83  7.1  33.0  33.0

       ...      wSL/C (pi)  wXX/C (pi)  O-Swing% (pi)  Z-Swing% (pi)
336    ...            1.76       22.85          0.364          0.665
236    ...            2.62         NaN          0.371          0.670
472    ...            3.92         NaN          0.336          0.598
235    ...            2.42         NaN          0.329          0.618
256    ...            0.74         NaN          0.339          0.635

     Swing% (pi)  O-Contact% (pi)  Z-Contact% (pi)  Contact% (pi)  Zone% (pi)
336        0.511            0.478            0.811          0.689       0.487
236        0.525            0.536            0.831          0.730       0.515
472        0.468            0.485            0.886          0.744       0.505
235        0.468            0.595            0.856          0.762       0.483
256        0.484            0.563            0.873          0.763       0.492

     Pace (pi)
336       23.4
236       23.7
472       24.6
235       23.3
256       23.4
```

Other Interesting Dependencies

schedule_and_record

```
>>> from pybaseball import schedule_and_record
>>> data = schedule_and_record(1927, 'NYY')
>>> data.head()
              Date   Tm Home_Away  Opp W/L     R   RA    Inn  W-L  Rank  \
1   Tuesday, Apr 12  NYY      Home  PHA   W   8.0  3.0   9.0  1-0   1.0
2  Wednesday, Apr 13  NYY      Home  PHA   W  10.0  4.0   9.0  2-0   1.0
3   Thursday, Apr 14  NYY      Home  PHA   T   9.0  9.0  10.0  2-0   1.0
4     Friday, Apr 15  NYY      Home  PHA   W   6.0  3.0   9.0  3-0   1.0
5   Saturday, Apr 16  NYY      Home  BOS   W   5.0  2.0   9.0  4-0   1.0

       GB      Win     Loss  Save  Time D/N  Attendance  Streak
1   Tied     Hoyt    Grove  None  2:05   D     72000.0       1
2  up 0.5  Ruether     Gray  None  2:15   D      8000.0       2
3   Tied     None     None  None  2:50   D      9000.0       2
4   Tied  Pennock    Ehmke  None  2:27   D     16000.0       3
5  up 1.0  Shocker  Ruffing  None  2:05   D     25000.0       4
```

standings

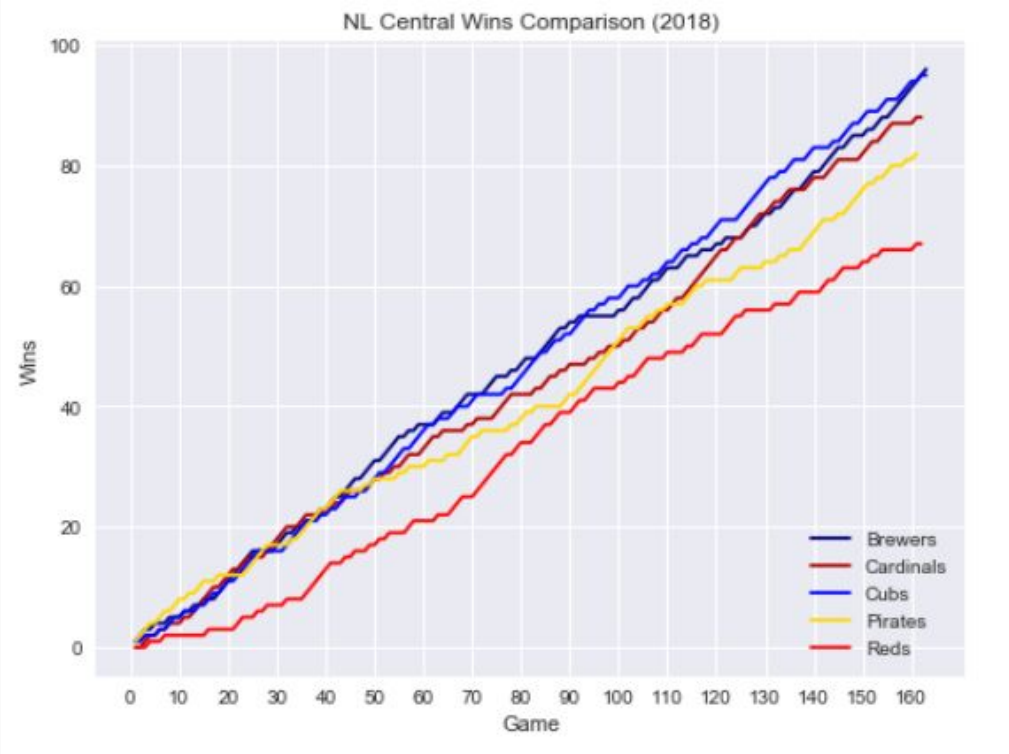
```
>>> from pybaseball import standings
>>> data = standings(2016)[4]
>>> print(data)
                    Tm    W   L  W-L%    GB
1          Chicago Cubs  103  58  .640    --
2  St. Louis Cardinals   86  76  .531  17.5
3   Pittsburgh Pirates   78  83  .484  25.0
4    Milwaukee Brewers   73  89  .451  30.5
5       Cincinnati Reds  68  94  .420  35.5
```

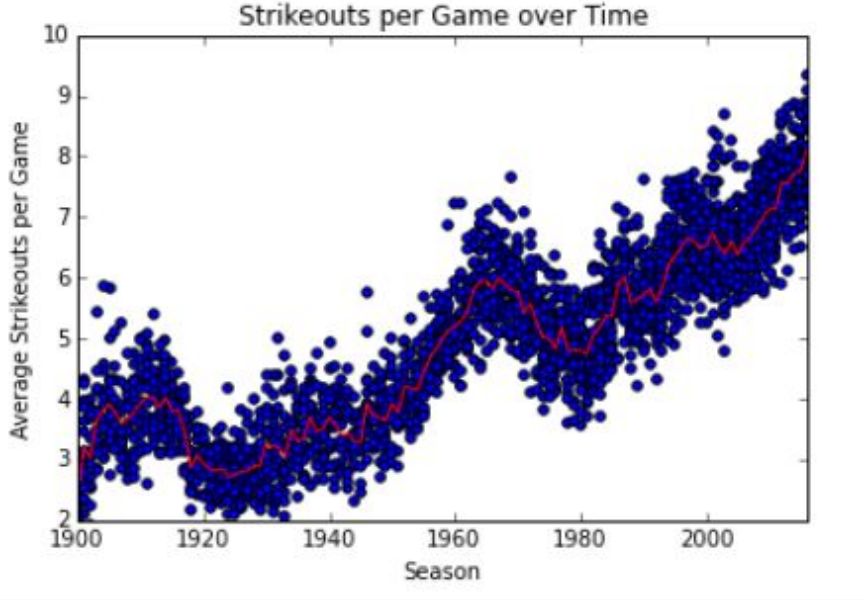
batting_stats_range

```
>>> from pybaseball import batting_stats_range
>>> data = batting_stats_range('2017-05-01', '2017-05-08')
>>> data.head()
          Name  Age  #days     Lev          Tm  G  PA  AB  R  H  ...  HBP
1   Jose Abreu   30     69  MLB-AL     Chicago  7  31  30  5  9  ...    0
2   Lane Adams   27     69  MLB-NL     Atlanta  6   6   6  0  2  ...    0
3   Matt Adams   28     68  MLB-NL   St. Louis  6   9   9  2  4  ...    0
4   Jim Adduci   32     69  MLB-AL     Detroit  6  24  21  3  5  ...    0
5  Tim Adleman   29     72  MLB-NL  Cincinnati  1   2   2  0  0  ...    0

   SH  SF  GDP  SB  CS     BA    OBP    SLG    OPS
1   0   0    1   0   0  0.300  0.323  0.667  0.989
2   0   0    1   1   0  0.333  0.333  0.333  0.667
3   0   0    0   0   0  0.444  0.444  0.778  1.222
4   0   0    0   0   0  0.238  0.333  0.381  0.714
5   0   0    0   0   0  0.000  0.000  0.000  0.000
```

Potential Experiments





Future Experiments

There is a near limitless amount of experiments we could perform using this data. The purpose of the project is for users to pull necessary and perform analysis on their own. From analyzing pitch velocity, on base percentage, slugging, earned run average, there is a plethora of statistics to run tests on. I will come up with more specifics as the project progresses.