Butts in Seats

A Data-Driven Look at Major League Baseball Attendance

Paul Giesting, PhD
Metis New York DS23

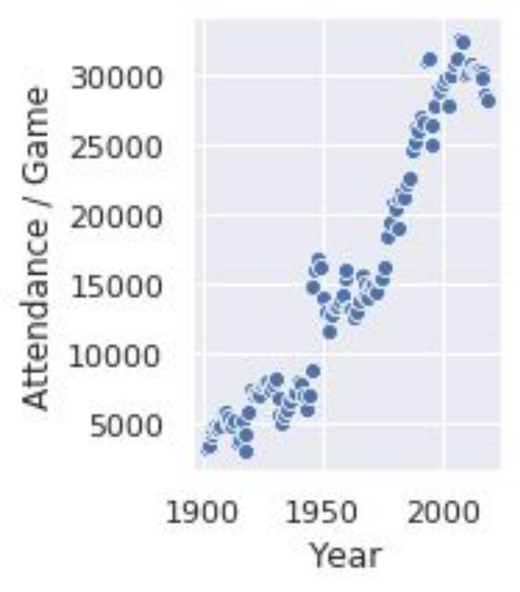
Baseball Attendance in the 21st Century

The Problem

After a century of increase, game attendance is slipping.

This is partly a side effect of different stadium and revenue design, but let's consider whether the game itself is contributing.

The Data



The Approach

Baseball has more things to count than any other sport, and a longer history to boot.

Let's put that data to work.

Workflow

Import & parse

Get the data imported and converted to a usable form.

Visualize & judge

Not every stat can be used or is equally useful.

After inspection, dead ball era data is [sigh] not valuable for this project.

A curated set of statistics were evaluated and cut down to create a model.

Model & ponder

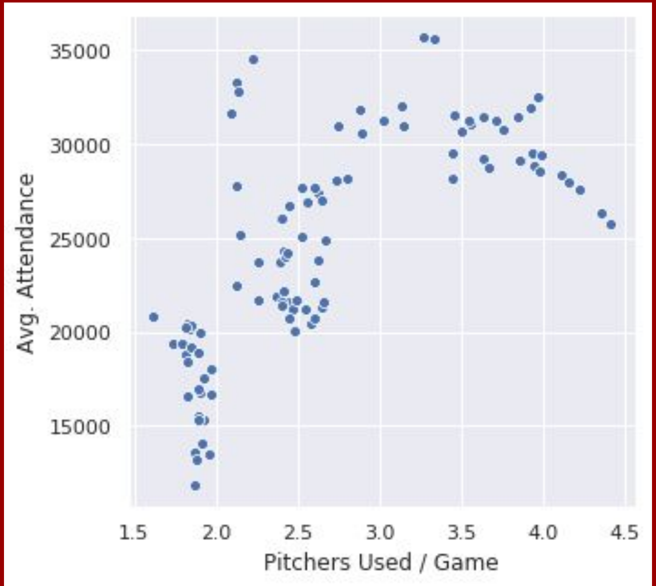
LassoCV: repeated train-test splits, checking the number of times different features are removed.

Statsmodels: feature *p*-values, kicking the worst stats off the island, iterate to stabilization.

The Insights

Strategy

Reinforcing Conclusions

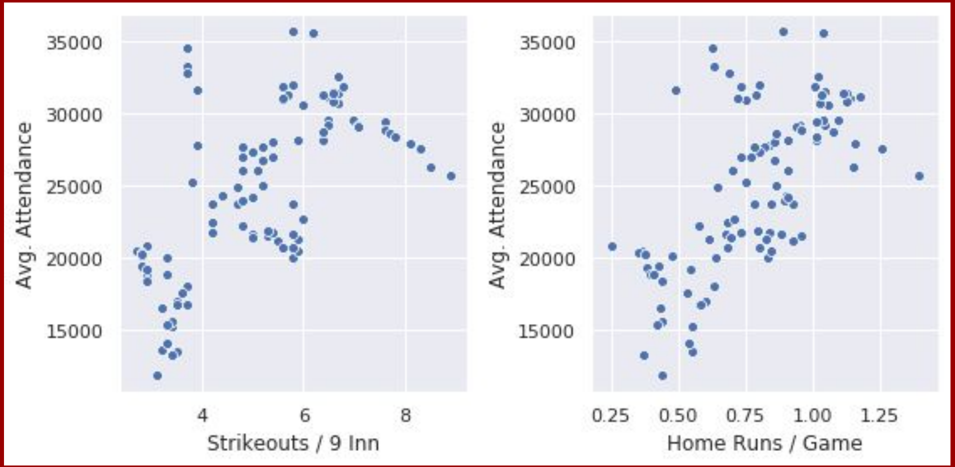


Pitching changes:

- The game has passed a tipping point.

Strategy

Usable Insights

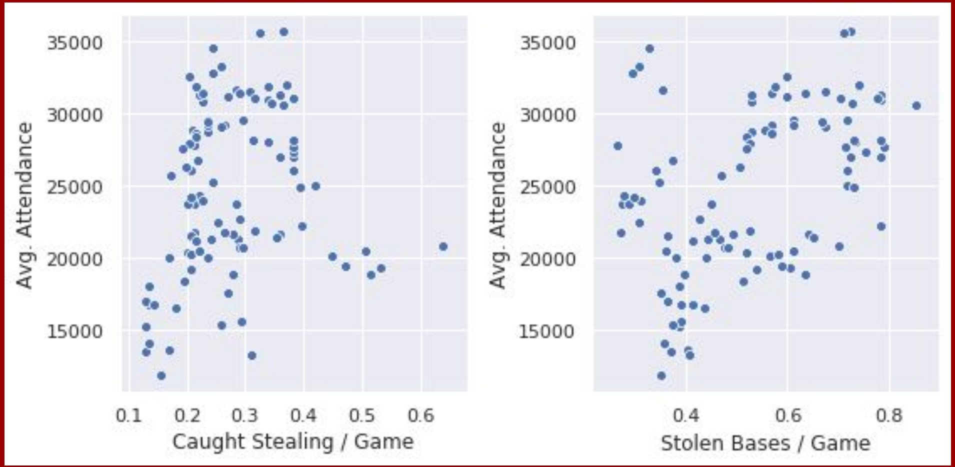


Strikeouts and home runs:

- Do fans want to wait around for the Three True Outcomes?
- Idea: Reverse the trend of short outfield fences.

Strategy

Rethinking Culture



Stolen bases:

- Base stealers *and* catchers are heroes to the fans.
- Culture / scouting changes.

The game has passed a tipping point. Too much of two good things.

HR / SO

Pitchers / Game

This is also past a tipping point. Intervention was wise.

Steals are exciting, even… especially… when a team’s favorite catcher makes the play.

Caught Stealing

Fielding

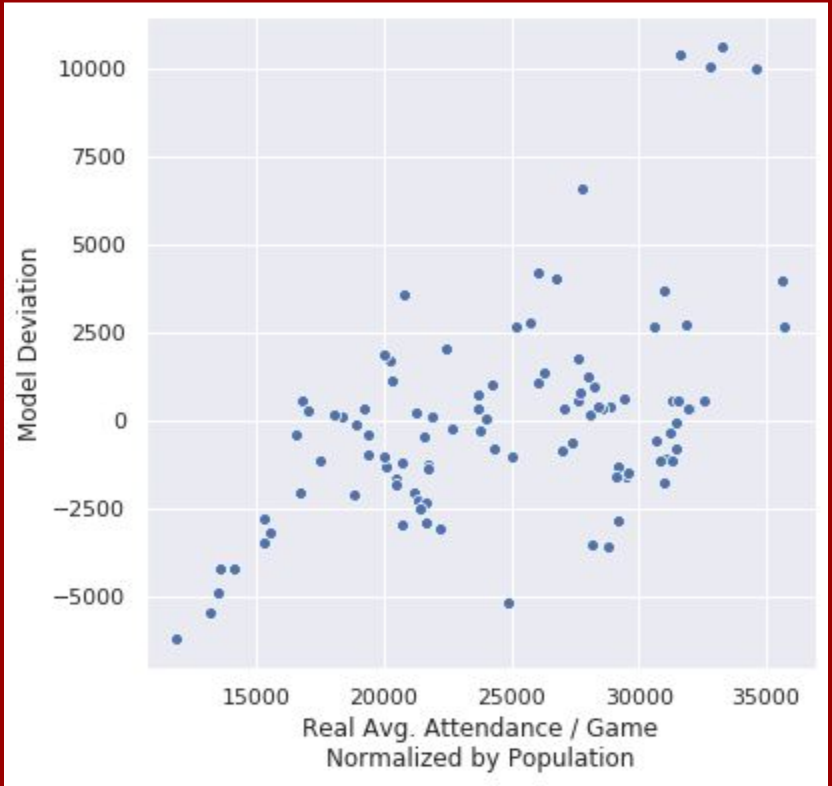
Today’s fielders have been perfecting their craft for a lifetime, and the fans appreciate it.

Bunts

The data suggest that fans want action and visible strategy, even if strategically it’s unwise.

Appendix

Model Quality



Main Model

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.724
Model:                            OLS   Adj. R-squared:                  0.712
Method:                 Least Squares   F-statistic:                     62.33
Date:                Thu, 16 Apr 2020   Prob (F-statistic):           9.67e-26
Time:                        23:43:22   Log-Likelihood:                -372.88
No. Observations:                 100   AIC:                             755.8
Df Residuals:                      95   BIC:                             768.8
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       -115.6298     18.879     -6.125      0.000    -153.109     -78.151
SO9          -12.9784      2.486     -5.220      0.000     -17.915      -8.042
Pitchers/G   145.4968     16.835      8.643      0.000     112.075     178.918
PitG^2       -17.4525      2.382     -7.327      0.000     -22.181     -12.724
CS/G          32.9747     11.316      2.914      0.004      10.510      55.439
==============================================================================
Omnibus:                       35.735   Durbin-Watson:                   0.551
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               77.947
Skew:                           1.367   Prob(JB):                     1.19e-17
Kurtosis:                       6.351   Cond. No.                         263.
==============================================================================
```

Alternative Model with HR, Bunts, Fld%

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.763
Model:                            OLS   Adj. R-squared:                  0.748
Method:                 Least Squares   F-statistic:                     49.92
Date:                Thu, 16 Apr 2020   Prob (F-statistic):           5.63e-27
Time:                        19:36:08   Log-Likelihood:                -365.26
No. Observations:                 100   AIC:                             744.5
Df Residuals:                      93   BIC:                             762.8
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -1907.1394    364.545     -5.232      0.000   -2631.054   -1183.225
HR/G          29.6711      9.542      3.109      0.002      10.722      48.620
Pitchers/G    16.0803      5.944      2.705      0.008       4.277      27.884
SO9^2         -1.2423      0.215     -5.773      0.000      -1.670      -0.815
SH^2          12.8740      5.030      2.560      0.012       2.886      22.862
Fld^2       2033.2049    392.044      5.186      0.000    1254.682    2811.727
CS/G          48.6562     11.090      4.387      0.000      26.634      70.679
==============================================================================
Omnibus:                       33.816   Durbin-Watson:                   0.747
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               81.257
Skew:                           1.234   Prob(JB):                     2.27e-18
Kurtosis:                       6.662   Cond. No.                     1.83e+04
==============================================================================
```

Outlier Years

Alternative Model with HR, Bunts, Fld%

	Att	AttP	Diff
1933	39.544513	58.736799	-19.192286
1943	44.098289	69.470340	-25.372051
1946	105.481293	70.252783	35.228510
1947	110.934573	86.323748	24.610825
1948	115.337925	79.269745	36.068180
1949	109.270058	83.608718	25.661340
1981	82.978167	97.883077	-14.904910
1988	103.218814	91.624403	11.594411
1989	106.142128	89.182330	16.959798
1993	119.128963	102.510133	16.618830
1994	118.785391	103.667513	15.117878
2003	95.932577	110.367523	-14.434946

Predictive Model: Calibrated on 20th century data.

Linear features only.
(A last minute addition to show where I wanted to go.)

OLS Regression Results

Dep. Variable:	Att	R-squared:	0.796
Model:	OLS	Adj. R-squared:	0.776
Method:	Least Squares	F-statistic:	40.61
Date:	Fri, 17 Apr 2020	Prob (F-statistic):	1.05e-22
Time:	09:14:58	Log-Likelihood:	-290.54
No. Observations:	81	AIC:	597.1
Df Residuals:	73	BIC:	616.2
Df Model:	7		
Covariance Type:	nonrobust		

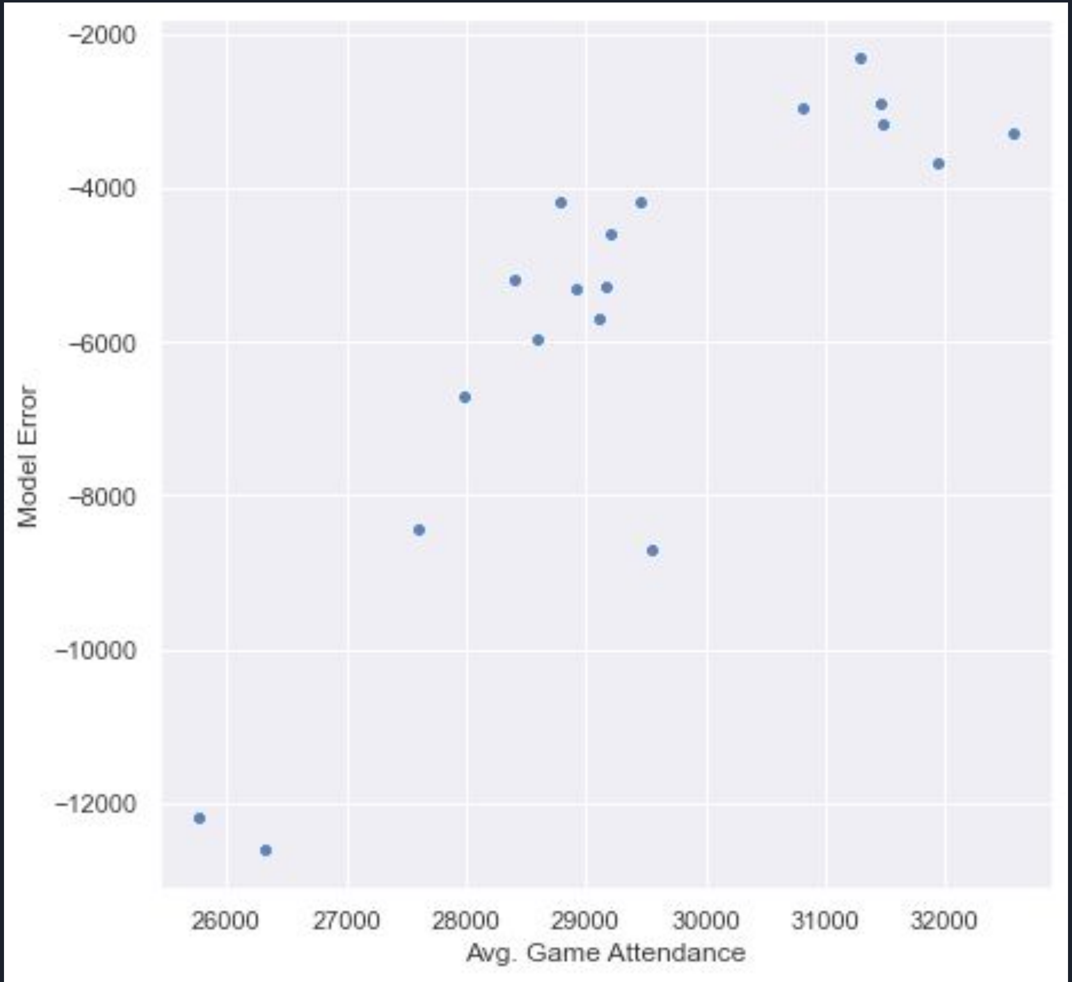
	coef	std err	t	P>|t|	[0.025	0.975]
const	-4832.6800	964.096	-5.013	0.000	-6754.120	-2911.240
Year	0.3959	0.177	2.238	0.028	0.043	0.748
SH/G	39.3055	11.103	3.540	0.001	17.176	61.435
SO9	-11.2172	2.655	-4.224	0.000	-16.509	-5.925
CS/G	94.0683	17.449	5.391	0.000	59.293	128.844
3B/G	188.7290	39.988	4.720	0.000	109.034	268.424
A/G	-39.1350	7.109	-5.505	0.000	-53.303	-24.967
Fld%	4634.9524	1065.883	4.348	0.000	2510.651	6759.253

Omnibus:	4.116	Durbin-Watson:	1.244
Prob(Omnibus):	0.128	Jarque-Bera (JB):	3.503
Skew:	0.498	Prob(JB):	0.174
Kurtosis:	3.212	Cond. No.	2.71e+06

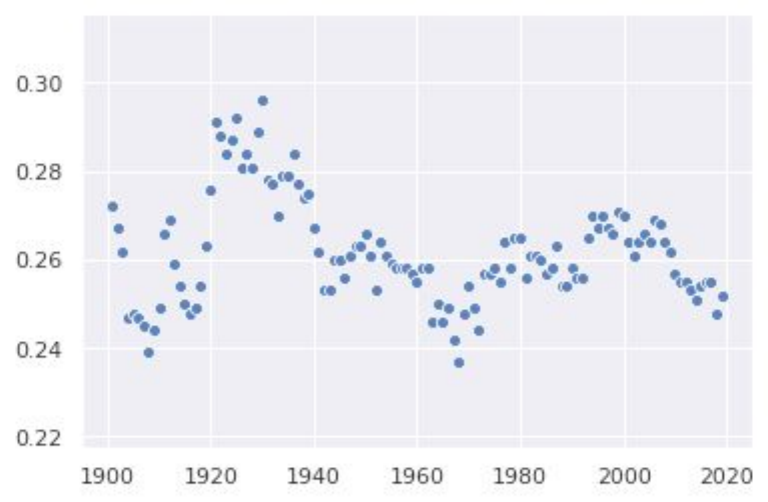
Predictive Model: Calibrated on 20th century data.

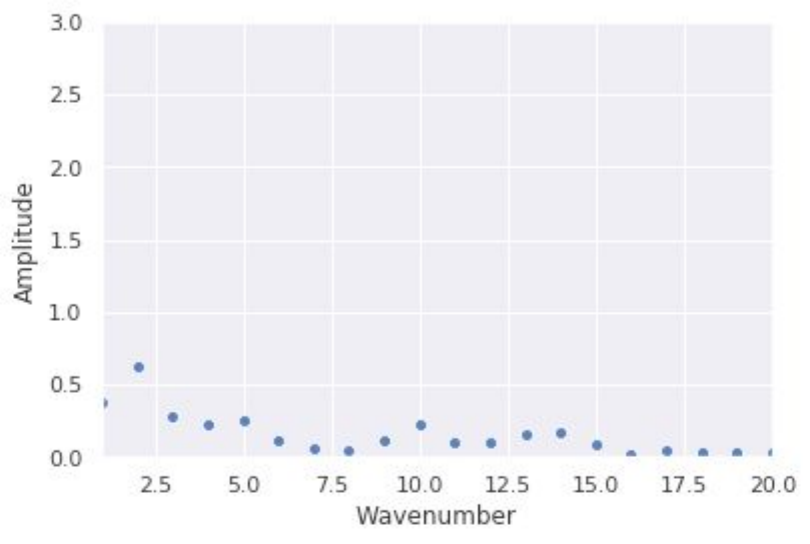
Systematically overpredicts 21st century data.

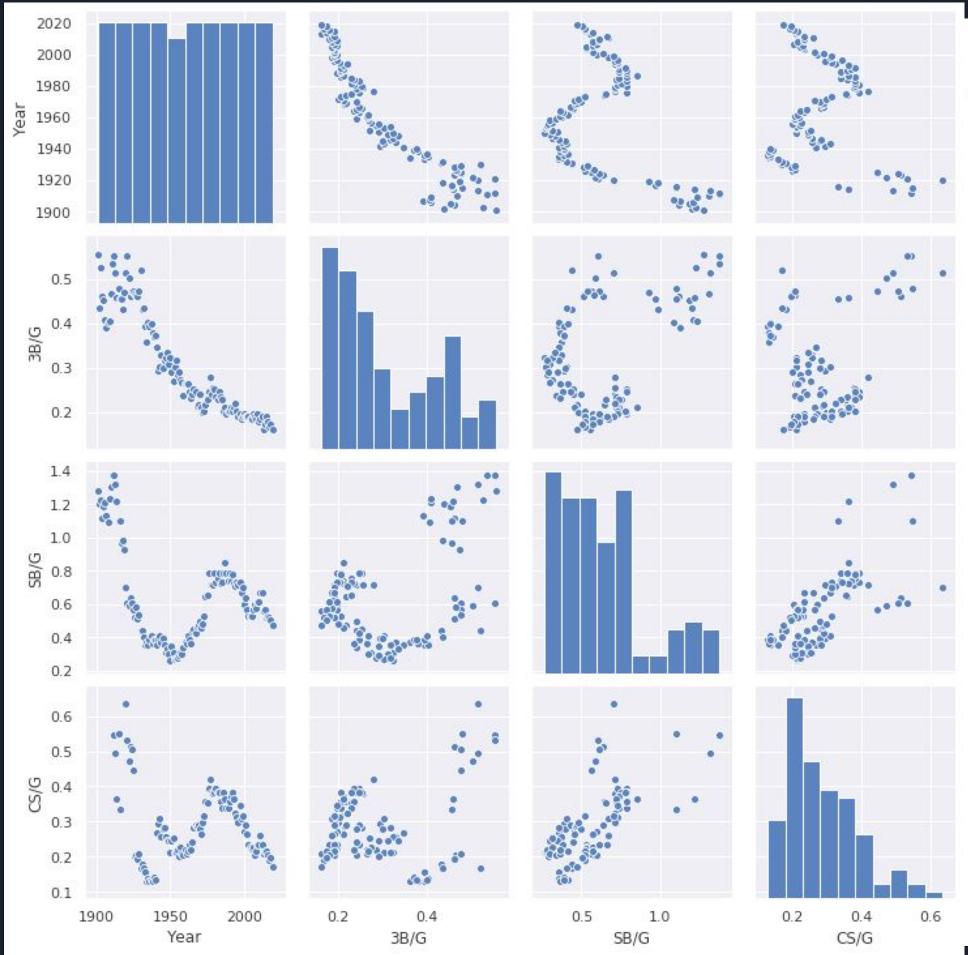
- Culture is shifting.
- Stadiums & revenue models are shifting.

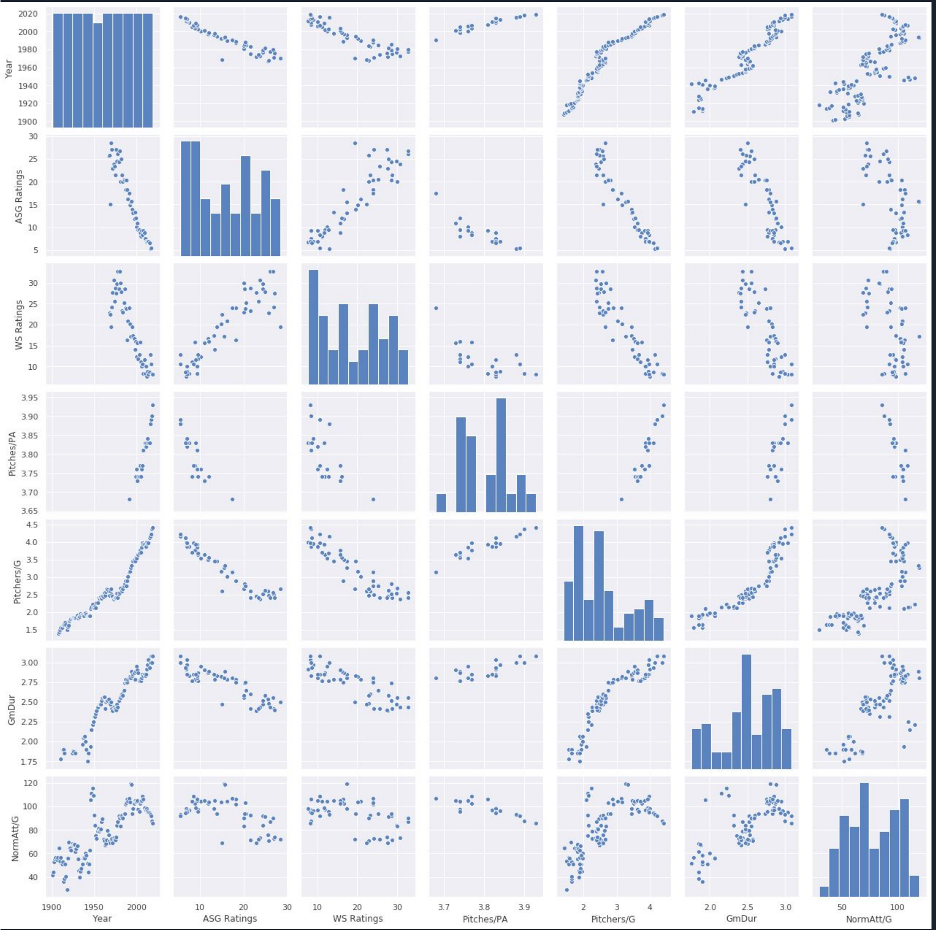


Periodicity: Batting Average






Year
3B/G
SB/G
CS/G


Year
ASG Ratings
WS Ratings
Pitches/PA
Pitchers/G
GmDur
NormAtt/G