The Effect of the Pitch Clock on MLB Players

ADTA 5340 & IPAC 4340 Team 1

Diana Bergeman

Eric Droegemeier

Gregory Ehlinger

Mauricio Sanchez

Triniti Lemmons

Major League Baseball (MLB), created in the late nineteenth century, is the United States' oldest professional sports league. The MLB is composed of thirty teams spread across the U.S. and Canada. During games, the objective of the pitcher is to throw three pitches across home plate into the strike zone. Any pitch thrown outside of the strike zone results in a ball. If four balls are thrown, the batter is granted the opportunity to walk to first base. The location of the strike zone differs from player to player, which forces the pitcher to constantly have to change the way in which they pitch the ball. The strike zone is the midpoint between the batter's shoulders and the top of their uniform pants. Since all players differ in height, the strike zone will be high for certain batters and lower for others. Pitchers must constantly be aware of their surroundings. They must know how many players of the opposite team are on base in the event that someone tries to steal a base. The pitcher's awareness is in addition to also focusing on the current batter and upcoming pitch.

The average regulation MLB game has nine innings, which lasts a total of approximately three hours long. Every season from 2016 to 2022 has lasted an average of no less than three hours in length. The 2021 season recorded the longest average game length in Major League Baseball history with a time of three hours and ten minutes (BETMGM, 2023). With the length of MLB games steadily increasing, league executives determined there was a need for change in order to optimize the baseball gaming experience for everyone involved.

The MLB is not unfamiliar with rule changes. Due to this, it was not surprising that the league introduced a new rule for the 2023 season. Beginning during the 2023 season, the MLB implemented a pitch clock with the goal of shortening the length of games. Following the conclusion of the previous play, this new timer gives pitchers 15 seconds to pitch the ball in the event that there are no batters on any of the bases. If there is a batter on base, then the pitcher is given 20 seconds to pitch the ball following the previous play. If the pitcher does not release the ball prior to the pitch clock expiring, the pitch will automatically be counted as a ball, which is in the favor of the batting team.

Our analysis seeks to answer the following question: Does the MLB pitch clock impact pitcher performance? The MLB as a league has a goal to improve the game experience. If the implementation hinders the abilities of pitchers to play to their full potential, then there is a chance they may have to re-evaluate the pitch clock's effectiveness at enhancing the game experience for all players, coaching staff, and spectators.

EDA/ Data Understanding/ Data Preparation

For our data sources, we took data from Fangraphs.com and from Statcast.com. Fangraphs sources data from Sports Info Solutions, which compiles data of every single pitch of every single MLB game, along with the pitch type, velocity, and result of the play. Statcast.com is an official website of Major League Baseball, which uses high speed cameras to collect data such as pitch spin rate, exit velocity, launch angle, etc. All stats were taken from Fangraphs, while Statcast was used as a reference only for the number of pitch count violations each individual pitcher had.

Our dataset consisted of pitcher data from 2022 and 2023. In 2022 alone, there were in total 708,540 pitches thrown in every major league baseball game. For our 2023 dataset, we took the total number of pitches thrown in the 2023 MLB season as of July 8th, 2023, which consisted of 413,067 pitches. Fangraphs does the data cleaning for us by compiling the results of every pitch and at-bat into statistics that can be used to assess pitching performance. We selected a mixture of statistics to analyze this, from regular counting stats to more advanced sabermetric stats. A list and explanation of each variable is found below.

Definition of variables

Explanations of statistics were taken from MLB or Fangraphs and sourced below. Our dataset consists of the following variables:

Name: The pitcher's first and last name.

Team: The team played for during the season. If the player played for multiple teams during the season, the data is missing.

IP: Innings Pitched. An inning consists of 3 outs. Partial innings are recorded in decimal form, with outs represented as .1 and .2. For example, Sandy Alcantara pitched 228.2 innings in 2022, which means he pitched 228 complete 3 out innings and got 2 of the 3 outs in the 229th inning. These are not mathematical decimals, because MLB Stats are not standard!

Pitches: Total number of pitches thrown.

Balls: Total number of balls thrown. An automatic ball is called on the pitcher for a pitch clock violation.

Strikes: Total number of strikes thrown. An automatic strike is called on the batter for a pitch clock violation.

ERA: Earned Run Average. The average number of runs a pitcher allows per game. Calculated as 9*ER/IP. An ER, earned run, is defined as any run that scores against a pitcher without the benefit of an error or a passed ball, according to MLB.com. If a pitcher exits the game with runners on base, any earned runs scored by those runners off of the reliever will count against him.

K% and BB%: These measure the percentage of plate appearances against the pitcher that resulted in a strikeout or a walk. K means strikeout while BB means base on balls, which is recorded when a walk has been issued.

HR/9: The number of home runs allowed by the pitcher per 9 innings pitched.

FIP: Fielding Independent Pitching. According to Piper Slowinski (2020), "FIP estimates the pitcher's run prevention independent of the performance of their defense. FIP is based on the outcomes that do not involve the defense: strikeouts, walks, hit by pitches, and home runs allowed. FIP uses those statistics and approximates a pitcher's ERA assuming average outcomes on balls in play." FIP is a more advanced statistic than ERA and uses league average coefficients based on changes in the leagues run environment, that is if the entire league is trending towards more offense (high scoring) or pitching and defense (low scoring.) The exact calculation is below:

$$FIP = (13*HR + 3*(BB+HBP) - 2*K)/IP + FIP\ Constant$$

xFIP: Expected FIP. According to Slowinski, "…is a statistic that estimates a pitcher's expected run prevention independent of the performance of their defense." In how xFIP differs from FIP, Slowinski writes "…it replaces a pitcher's home run total with an estimate of how many home runs they *should* have allowed given the number of fly balls they surrendered while assuming a league average home run to fly ball percentage." See the calculation for xFIP below:

$$xFIP = (13*(Fly\ Balls * LgHR/FB\%) + 3*(BB+HPB)-2*K)/IP + FIP\ Constant$$

ERA-: ERA-, sounded like ERA minus, is a pitcher's ERA in relation to the league average normalized to 100. Each point below or above 100 represents a percent above or below the league average. Lower is better. Ryne Stanek of the 2022 World Series Champion Houston Astros had an ERA- of 30, which means his ERA was 70% better than the average pitcher.

WPA: Win Probability Added shows the change in a team's win expectancy on a per plate appearance basis and credits (or debits) the player based on how much their action increased their team's chances of winning. WPA rewards players who perform in big moments of games, and you can think of WPA as being a statistic that measures "clutch." For historical context, the largest WPA on a single play in the World Series was 0.87, with Kirk Gibson's game-winning homer off Dennis Eckersley in game 1 of the 1988 World Series, which gave the Dodgers, who had been trailing 4-3 with 2 outs in the 9th, the victory.

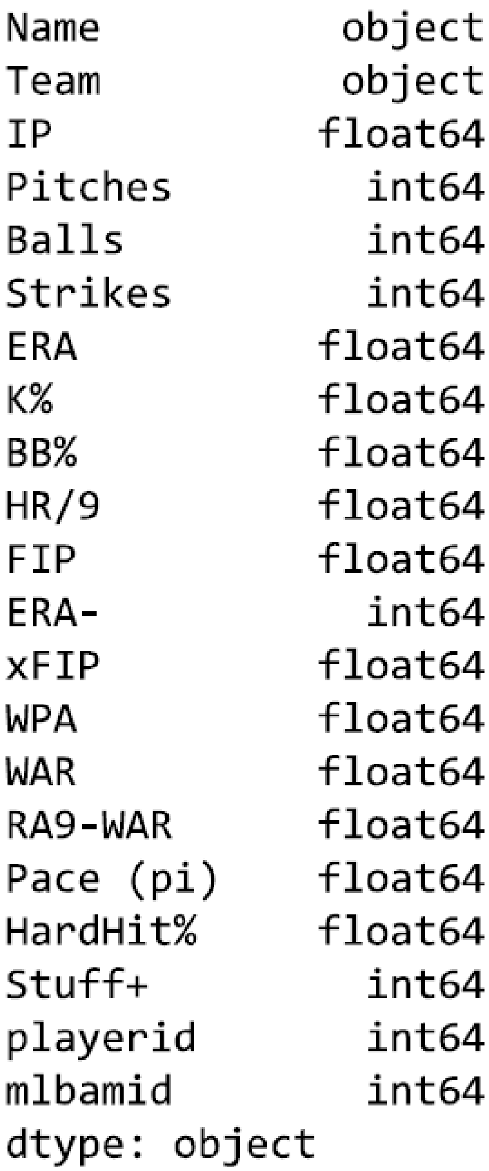
WAR: Wins Above Replacement. An attempt by Sabermetricians to summarize a player's total contributions to their team in one statistic. It looks to measure how many wins a player contributed to their team over a sub, or "replacement", player. The WAR leader among pitchers in 2022 was Aaron Nola, with 6.3 wins above a replacement player. The lowest was poor Adam Oller, who was calculated to have cost his team -1.1 WAR!

Pace(pi): The pace of play measure. This measures the average number of seconds between pitches for each individual pitcher. We should see a significant drop in pace by the slowest pitchers from 2022 to 2023 as the pitch clock was introduced to combat lengthy delays.

HardHit%: Represents a percentage of batted balls that resulted in hard contact by the batter. This is calculated based on hang time, location, trajectory, and exit velocity. The exact information is proprietary to Fangraphs and not public, but is included to see if there is a significant change because of the clock.

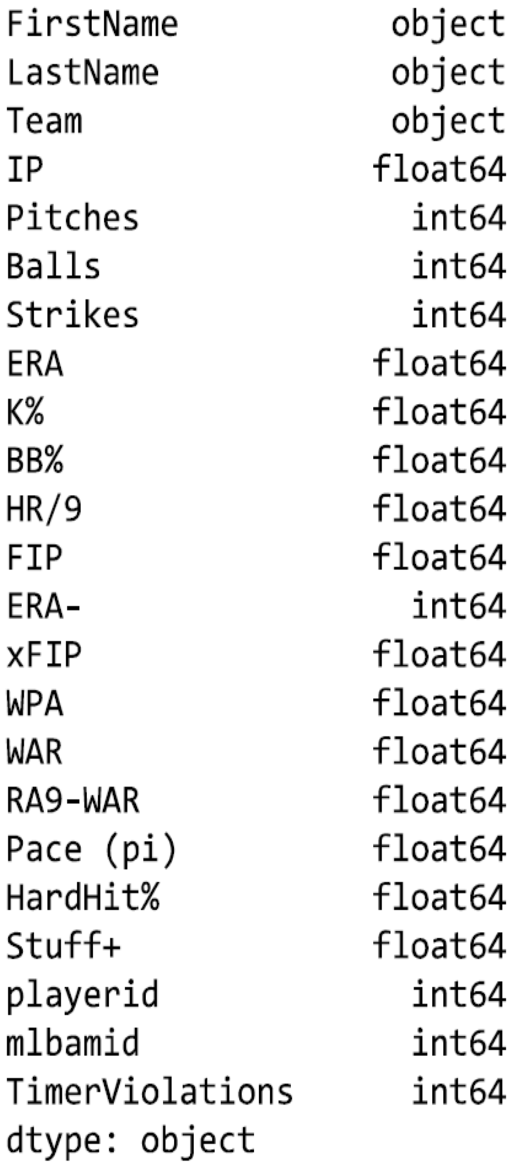
Stuff+: This is a newly invented measurement developed in 2023 (but extrapolated to previous seasons based on available data) to measure the physical characteristics of a pitch. According to Owen McGrattan, “Important features include, but are not limited to, release point, velocity, vertical and horizontal movement, and spin rate.) It measures the overall deception of a pitch in an attempt to explain why some pitches perform better than others at generating weak contact from the batter. Pitchers with a better Stuff+ tend to have more “deception” to their pitches. Major League hitters are talented enough to identify the spin of a pitch and be able to predict its location. Stuff+ includes an axis differential, which, in layman’s terms, is the observed movement from seam-shifted wake. According to Driveline, “Most pitch movement is explained by this force, with the movement coming perpendicular to the spin axis of the pitch.” Seam-shifted wake is the baseball terminology of the Magnus effect, which is an observable phenomenon of the movement of a spinning object moving through the air. “The Magnus effect is why soccer players can bend a soccer ball into the goal around a 5-person wall and why baseball pitchers can throw a breaking ball pitch.” (Seattle University, n.d.)

2022 Dataset



```
Name              object
Team              object
IP               float64
Pitches            int64
Balls              int64
Strikes            int64
ERA              float64
K%               float64
BB%              float64
HR/9             float64
FIP              float64
ERA-               int64
xFIP             float64
WPA              float64
WAR              float64
RA9-WAR          float64
Pace (pi)        float64
HardHit%         float64
Stuff+             int64
playerid           int64
mlbamid            int64
dtype: object
```

2023 Dataset



```
FirstName           object
LastName            object
Team                object
IP                 float64
Pitches              int64
Balls                int64
Strikes              int64
ERA                float64
K%                 float64
BB%                float64
HR/9               float64
FIP                float64
ERA-                 int64
xFIP               float64
WPA                float64
WAR                float64
RA9-WAR            float64
Pace (pi)          float64
HardHit%           float64
Stuff+             float64
playerid             int64
mlbamid              int64
TimerViolations      int64
dtype: object
```

Fig. 1

Distribution of data

Since we are comparing datasets from 2022 and 2023, we have two datasets. With the implementation of the pitch clock in 2023, our 2023 dataset had an additional variable called "Timer Violations" (*Fig. 1)*.

The graphical representation of the distribution of data for the 2022 season for each variable is shown in *Fig 2*. For Pitches, or the total number of pitches thrown for the season, per pitcher, we see that the lower end of the data range, the first quartile (Q1), which represents the lowest 25% of the data, at 998 pitches, the Median (Q2), or 50% of our data distribution, the upper quartile (Q3), which represents the lowest 75% of the data, at 2180 pitches thrown, and the upper whisker, which is Q3 + 1.5, are all below 3274 pitches thrown. The lowest value in the box plot, Q1-1.5, is 749 pitches for the season. We see similar distribution between highly correlated variables such as Balls, and Strikes.

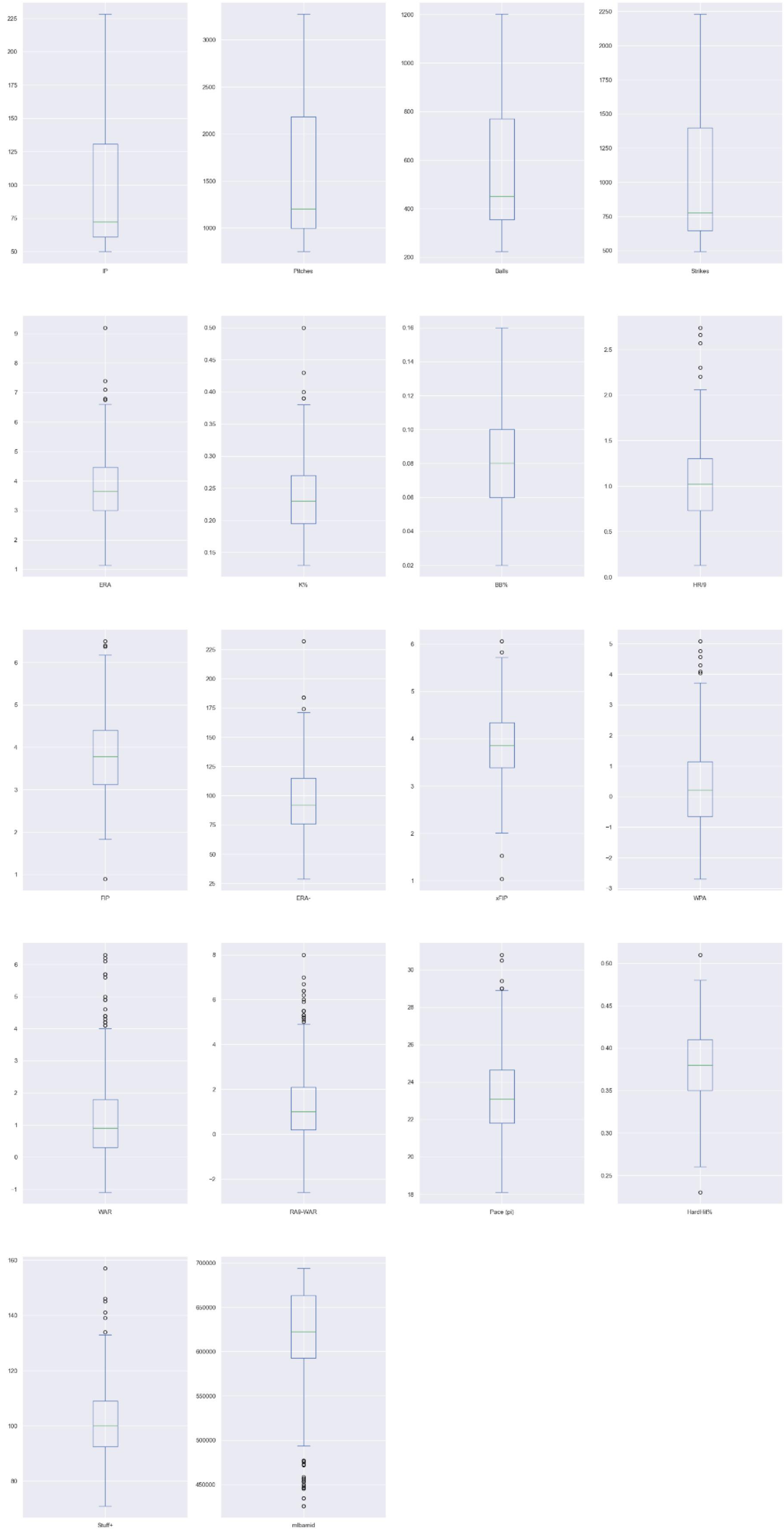

225
200
175
150
125
100
75
50
IP
3000
2500
2000
1500
1000
Pitches
1200
1000
800
600
400
200
Balls
2250
2000
1750
1500
1250
1000
750
500
Strikes
9
8
7
6
5
4
3
2
1
ERA
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
K%
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
BB%
2.5
2.0
1.5
1.0
0.5
0.0
HR/9
6
5
4
3
2
1
FIP
225
200
175
150
125
100
75
50
25
ERA-
6
5
4
3
2
1
xFIP
5
4
3
2
1
0
−1
−2
-3
WPA
6
5
4
3
2
1
0
−1
WAR
8
6
4
2
0
−2
RA9-WAR
30
28
26
24
22
20
18
Pace (pi)
0.50
0.45
0.40
0.35
0.30
0.25
HardHit%
160
140
120
100
80
Stuff+
700000
650000
600000
550000
500000
450000
mlbamid

Fig. 2

The graphical representation of the distribution of data for the 2023 season for each variable is shown in *Fig 3*. For Pitches, or the total number of pitches thrown for the season, per pitcher, we see that the lower end of the data range, the first quartile (Q1), which represents the lowest 25% of the data, at 137 pitches, the Median (Q2), or 50% of our data distribution, which is at 528 pitches, the upper quartile (Q3), which represents 75% of the data, at 697 pitches thrown, and the upper whisker, which is Q3 + 1.5, are all below 1,906 pitches thrown. The lowest value in the box plot, Q1-1.5, is 4 pitches for the season. Once again, we see similar distribution between highly correlated variables such as Balls, and Strikes. In the 2023 season, for Timer Violations, we see the majority of our data is distributed under 3 violations, with 3 or more violations considered outliers.

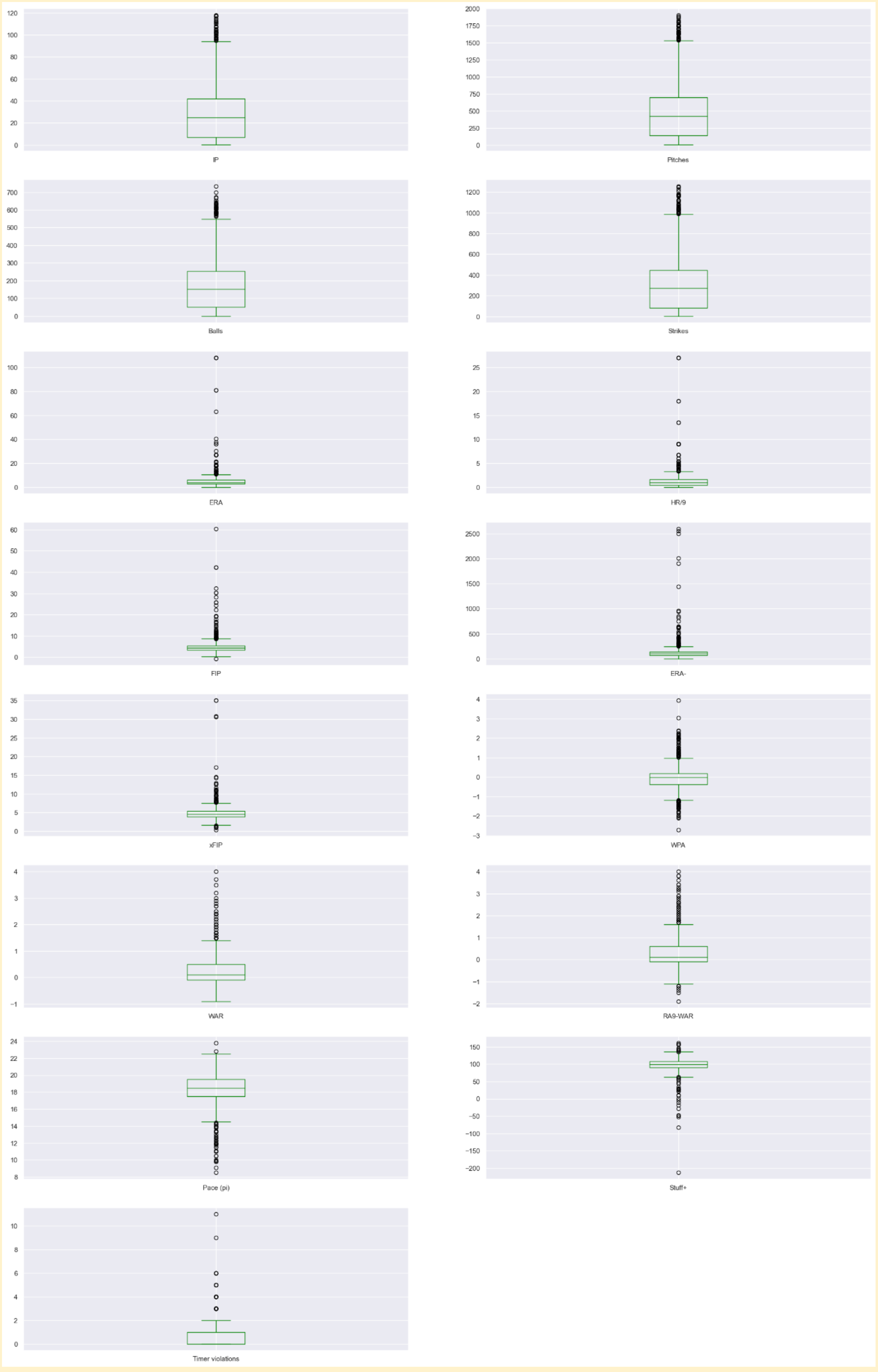

120
100
80
60
40
20
0
IP
2000
1750
1500
1250
1000
750
500
250
0
Pitches
700
600
500
400
300
200
100
0
Balls
1200
1000
800
600
400
200
0
Strikes
100
80
60
40
20
0
ERA
25
20
15
10
5
0
HR/9
60
50
40
30
20
10
0
FIP
2500
2000
1500
1000
500
0
ERA-
35
30
25
20
15
10
5
0
xFIP
4
3
2
1
0
−1
−2
−3
WPA
4
3
2
1
0
−1
WAR
4
3
2
1
0
−1
−2
RA9-WAR
24
22
20
18
16
14
12
10
8
Pace (pi)
150
100
50
0
−50
−100
−150
−200
Stuff+
10
8
6
4
2
0
Timer violations

Fig. 3

The 20023 data for many of the other variables, including Innings Pitched (IP), Pitches, Balls, and Strikes are positively skewed, which is parallel to the visualization for the number of outliers we have for those variables as shown in *Fig. 3*. The data for Pace(pi) and Stuff+ is negatively skewed, meaning that we have outliers in the lower ranges for that variable, and again parallels the visualization from *Fig. 3*.

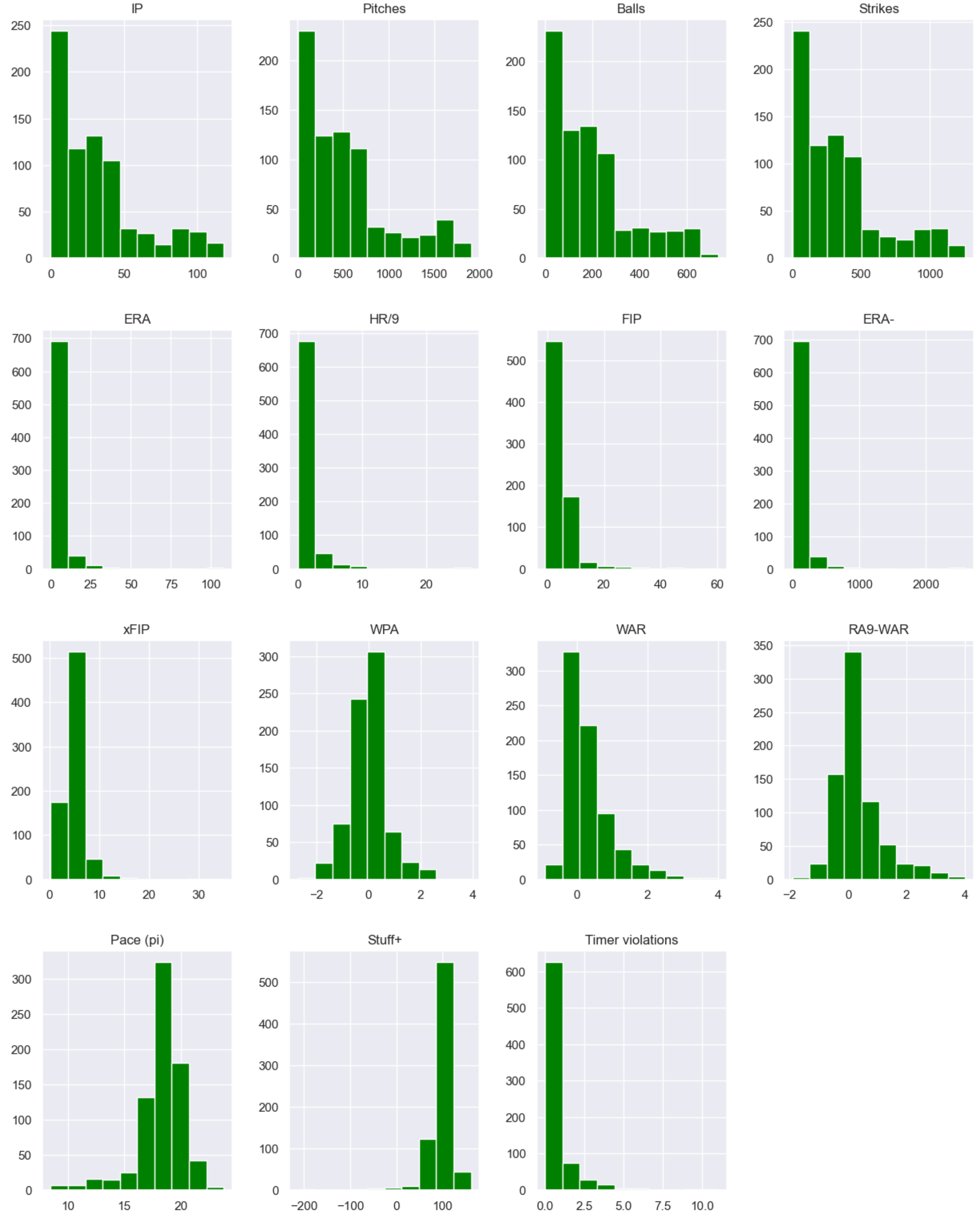
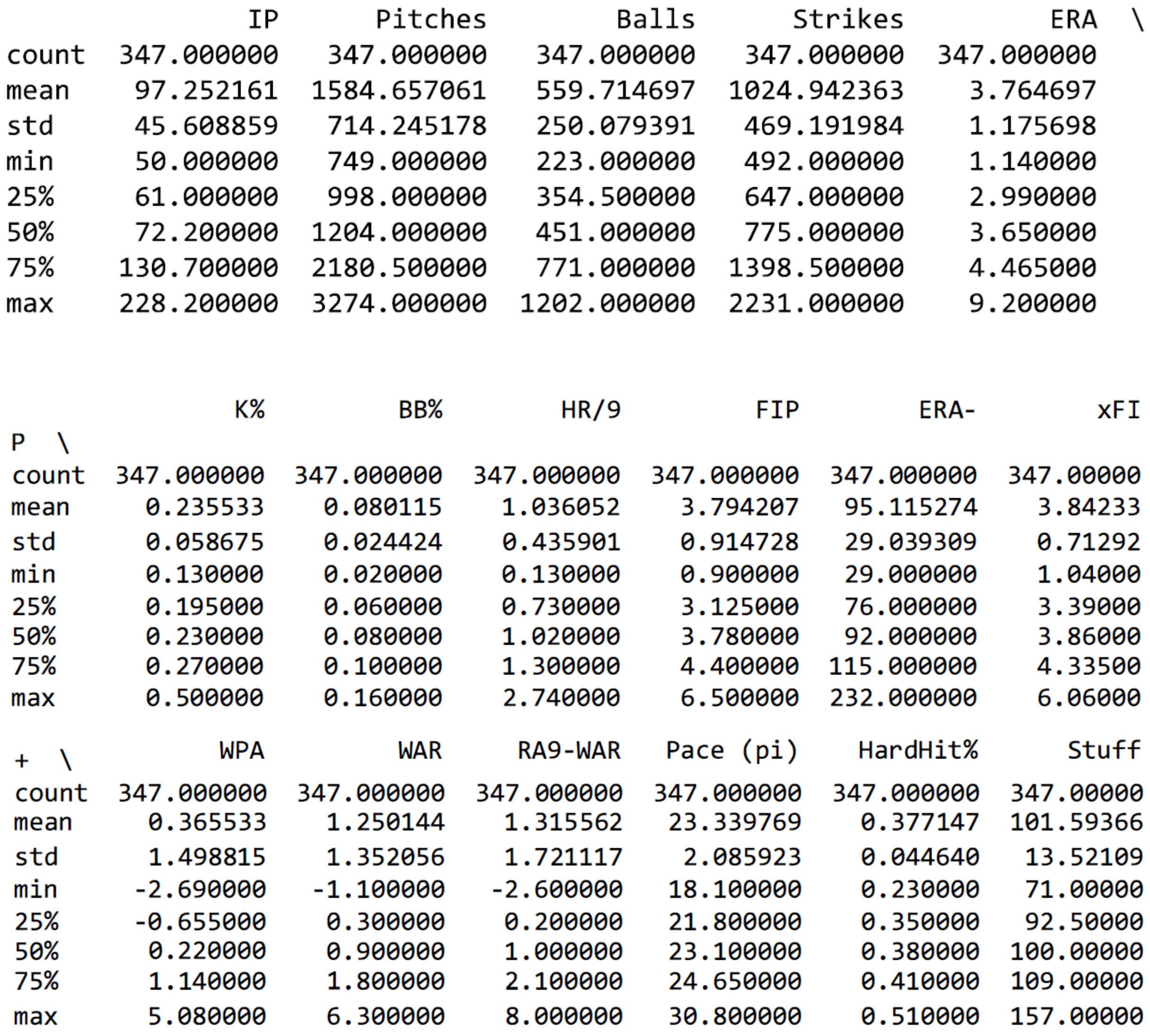


Fig. 4

Summary statistics

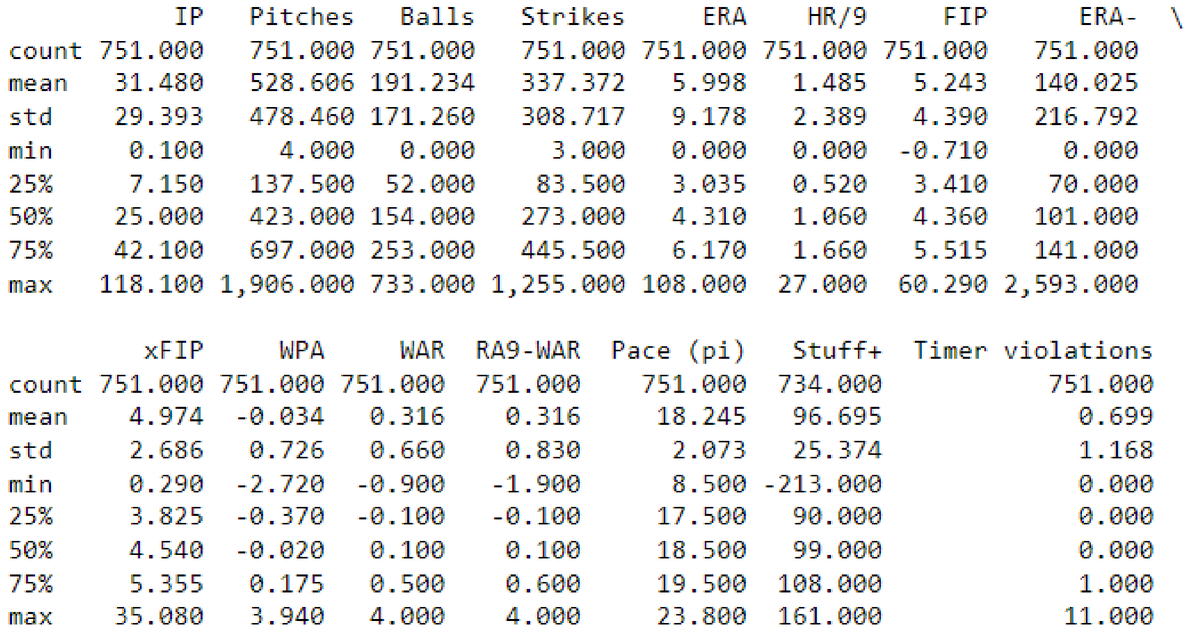
The average of all data points in our dataset for 2022, the central tendency of data, for Pitches is 1,584.66 (*Fig. 5),* 559.71 for Balls, and 1,204.94 for Strikes. Because the mean represents the average of all data points, this statistic is sensitive to the outliers in these variables. The Standard Deviation (std) indicates the variability of the data; with a std of 714.25 for Pitches, we see very high variability.



```
                IP      Pitches        Balls      Strikes         ERA  \
count   347.000000   347.000000   347.000000   347.000000  347.000000
mean     97.252161  1584.657061   559.714697  1024.942363    3.764697
std      45.608859   714.245178   250.079391   469.191984    1.175698
min      50.000000   749.000000   223.000000   492.000000    1.140000
25%      61.000000   998.000000   354.500000   647.000000    2.990000
50%      72.200000  1204.000000   451.000000   775.000000    3.650000
75%     130.700000  2180.500000   771.000000  1398.500000    4.465000
max     228.200000  3274.000000  1202.000000  2231.000000    9.200000


               K%         BB%        HR/9         FIP        ERA-        xFI
P  \
count  347.000000  347.000000  347.000000  347.000000  347.000000  347.00000
mean     0.235533    0.080115    1.036052    3.794207   95.115274    3.84233
std      0.058675    0.024424    0.435901    0.914728   29.039309    0.71292
min      0.130000    0.020000    0.130000    0.900000   29.000000    1.04000
25%      0.195000    0.060000    0.730000    3.125000   76.000000    3.39000
50%      0.230000    0.080000    1.020000    3.780000   92.000000    3.86000
75%      0.270000    0.100000    1.300000    4.400000  115.000000    4.33500
max      0.500000    0.160000    2.740000    6.500000  232.000000    6.06000

+  \          WPA         WAR     RA9-WAR   Pace (pi)    HardHit%      Stuff
count  347.000000  347.000000  347.000000  347.000000  347.000000  347.00000
mean     0.365533    1.250144    1.315562   23.339769    0.377147  101.59366
std      1.498815    1.352056    1.721117    2.085923    0.044640   13.52109
min     -2.690000   -1.100000   -2.600000   18.100000    0.230000   71.00000
25%     -0.655000    0.300000    0.200000   21.800000    0.350000   92.50000
50%      0.220000    0.900000    1.000000   23.100000    0.380000  100.00000
75%      1.140000    1.800000    2.100000   24.650000    0.410000  109.00000
max      5.080000    6.300000    8.000000   30.800000    0.510000  157.00000
```

Fig. 5

For the 2023 data, the average of all data points in our dataset, the central tendency of data, for Timer Violations is .699 and the average for Pitches is 528.61 (*Fig. 6)*. Because the mean represents the average of all data points, this statistic is sensitive to the outliers in these variables. With a Standard Deviation (std) of 1.168 for Timer Violations we have low variability in the data but with a std of 478.46 for Pitches, we see high variability.



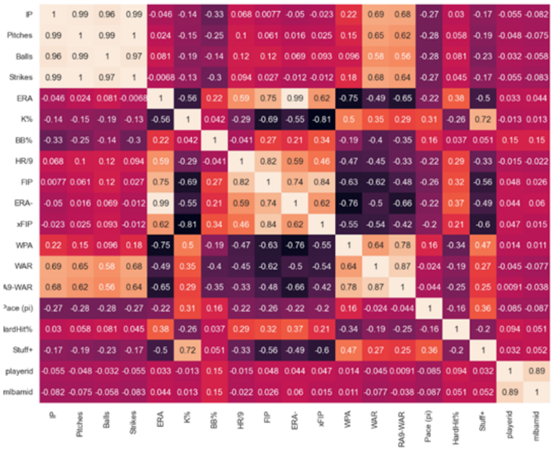
```
            IP    Pitches   Balls    Strikes      ERA     HR/9      FIP       ERA-  \
count  751.000    751.000 751.000    751.000  751.000  751.000  751.000    751.000
mean    31.480    528.606 191.234    337.372    5.998    1.485    5.243    140.025
std     29.393    478.460 171.260    308.717    9.178    2.389    4.390    216.792
min      0.100      4.000   0.000      3.000    0.000    0.000   -0.710      0.000
25%      7.150    137.500  52.000     83.500    3.035    0.520    3.410     70.000
50%     25.000    423.000 154.000    273.000    4.310    1.060    4.360    101.000
75%     42.100    697.000 253.000    445.500    6.170    1.660    5.515    141.000
max    118.100  1,906.000 733.000  1,255.000  108.000   27.000   60.290  2,593.000

          xFIP      WPA     WAR  RA9-WAR  Pace (pi)   Stuff+  Timer violations
count  751.000  751.000 751.000  751.000    751.000  734.000           751.000
mean     4.974   -0.034   0.316    0.316     18.245   96.695             0.699
std      2.686    0.726   0.660    0.830      2.073   25.374             1.168
min      0.290   -2.720  -0.900   -1.900      8.500 -213.000             0.000
25%      3.825   -0.370  -0.100   -0.100     17.500   90.000             0.000
50%      4.540   -0.020   0.100    0.100     18.500   99.000             0.000
75%      5.355    0.175   0.500    0.600     19.500  108.000             1.000
max     35.080    3.940   4.000    4.000     23.800  161.000            11.000
```

Fig. 6

Correlation between Variables

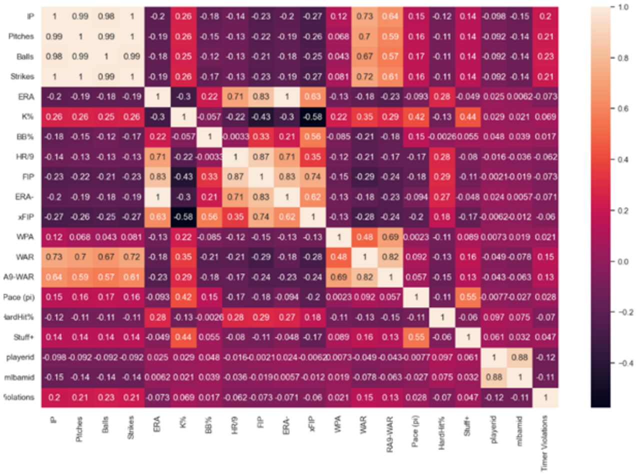
When comparing the correlation between key variables in the 2022 and 2023 data, we see our first indication of the answer to our research question. Below, on the left, for 2022, (*fig. 7)* we see higher correlation between Innings Pitched(ip) and pitches, balls, strikes, which makes sense; the higher number of innings a pitcher participates in, the higher number of pitches are thrown, and the higher the ball count and strike count climb.

With the assumption that the pitch clock and threat of a timer violation is going to impact athlete performance, we'd expect to see a high correlation in the 2023 data, below right (*Fig 7.)* between Timer Violations and Balls and Strikes, however those variables have relatively low correlation, with .23 and .21 respectively.



	IP	Pitches	Balls	Strikes	ERA	K%	BB%	HR/9	FIP	ERA-	xFIP	WPA	WAR	RA9-WAR	Pace (pi)	HardHit%	Stuff+	playerid	mlbamid
IP	1	0.99	0.96	0.99	-0.046	-0.14	-0.33	0.068	0.0077	-0.05	-0.023	0.22	0.69	0.68	-0.27	0.03	-0.17	-0.055	-0.082
Pitches	0.99	1	0.99	1	0.024	-0.15	-0.25	0.1	0.061	0.016	0.025	0.15	0.65	0.62	-0.28	0.058	-0.19	-0.048	-0.075
Balls	0.96	0.99	1	0.97	0.081	-0.19	-0.14	0.12	0.12	0.069	0.093	0.096	0.58	0.56	-0.28	0.081	-0.23	-0.032	-0.058
Strikes	0.99	1	0.97	1	-0.0068	-0.13	-0.3	0.094	0.027	-0.012	-0.012	0.18	0.68	0.64	-0.27	0.045	-0.17	-0.055	-0.083
ERA	-0.046	0.024	0.081	-0.0068	1	-0.56	0.22	0.59	0.75	0.99	0.62	-0.75	-0.49	-0.65	-0.22	0.38	-0.5	0.033	0.044
K%	-0.14	-0.15	-0.19	-0.13	-0.56	1	0.042	-0.29	-0.69	-0.55	-0.81	0.5	0.35	0.29	0.31	-0.26	0.72	-0.013	0.013
BB%	-0.33	-0.25	-0.14	-0.3	0.22	0.042	1	-0.041	0.27	0.21	0.34	-0.19	-0.4	-0.35	0.16	0.037	0.051	0.15	0.15
HR/9	0.068	0.1	0.12	0.094	0.59	-0.29	-0.041	1	0.82	0.59	0.46	-0.47	-0.45	-0.33	-0.22	0.29	-0.33	-0.015	-0.022
FIP	0.0077	0.061	0.12	0.027	0.75	-0.69	0.27	0.82	1	0.74	0.84	-0.63	-0.62	-0.48	-0.26	0.32	-0.56	0.048	0.026
ERA-	-0.05	0.016	0.069	-0.012	0.99	-0.55	0.21	0.59	0.74	1	0.62	-0.76	-0.5	-0.66	-0.22	0.37	-0.49	0.044	0.06
xFIP	-0.023	0.025	0.093	-0.012	0.62	-0.81	0.34	0.46	0.84	0.62	1	-0.55	-0.54	-0.42	-0.2	0.21	-0.6	0.047	0.015
WPA	0.22	0.15	0.096	0.18	-0.75	0.5	-0.19	-0.47	-0.63	-0.76	-0.55	1	0.64	0.78	0.16	-0.34	0.47	0.014	0.011
WAR	0.69	0.65	0.58	0.68	-0.49	0.35	-0.4	-0.45	-0.62	-0.5	-0.54	0.64	1	0.87	-0.024	-0.19	0.27	-0.045	-0.077
RA9-WAR	0.68	0.62	0.56	0.64	-0.65	0.29	-0.35	-0.33	-0.48	-0.66	-0.42	0.78	0.87	1	-0.044	-0.25	0.25	0.0091	-0.038
Pace (pi)	-0.27	-0.28	-0.28	-0.27	-0.22	0.31	0.16	-0.22	-0.26	-0.22	-0.2	0.16	-0.024	-0.044	1	-0.16	0.36	-0.085	-0.087
HardHit%	0.03	0.058	0.081	0.045	0.38	-0.26	0.037	0.29	0.32	0.37	0.21	-0.34	-0.19	-0.25	-0.16	1	-0.2	0.094	0.051
Stuff+	-0.17	-0.19	-0.23	-0.17	-0.5	0.72	0.051	-0.33	-0.56	-0.49	-0.6	0.47	0.27	0.25	0.36	-0.2	1	0.032	0.052
playerid	-0.055	-0.048	-0.032	-0.055	0.033	-0.013	0.15	-0.015	0.048	0.044	0.047	0.014	-0.045	0.0091	-0.085	0.094	0.032	1	0.89
mlbamid	-0.082	-0.075	-0.058	-0.083	0.044	0.013	0.15	-0.022	0.026	0.06	0.015	0.011	-0.077	-0.038	-0.087	0.051	0.052	0.89	1

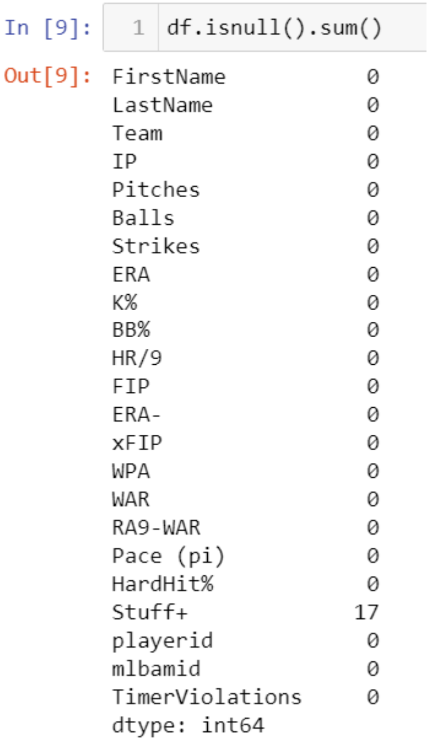
Fig. 7



	IP	Pitches	Balls	Strikes	ERA	K%	BB%	HR/9	FIP	ERA-	xFIP	WPA	WAR	RA9-WAR	Pace (pi)	HardHit%	Stuff+	playerid	mlbamid	Timer Violations
IP	1	0.99	0.98	1	-0.2	0.26	-0.18	-0.14	-0.23	-0.2	-0.27	0.12	0.73	0.64	0.15	-0.12	0.14	-0.098	-0.15	0.2
Pitches	0.99	1	0.99	1	-0.19	0.26	-0.15	-0.13	-0.22	-0.19	-0.26	0.068	0.7	0.59	0.16	-0.11	0.14	-0.092	-0.14	0.21
Balls	0.98	0.99	1	0.99	-0.18	0.25	-0.12	-0.13	-0.21	-0.18	-0.25	0.043	0.67	0.57	0.17	-0.11	0.14	-0.092	-0.14	0.23
Strikes	1	1	0.99	1	-0.19	0.26	-0.17	-0.13	-0.23	-0.19	-0.27	0.081	0.72	0.61	0.16	-0.11	0.14	-0.092	-0.14	0.21
ERA	-0.2	-0.19	-0.18	-0.19	1	-0.3	0.22	0.71	0.83	1	0.63	-0.13	-0.18	-0.23	-0.093	0.28	-0.049	0.025	0.0062	-0.073
K%	0.26	0.26	0.25	0.26	-0.3	1	-0.057	-0.22	-0.43	-0.3	-0.58	0.22	0.35	0.29	0.42	-0.13	0.44	0.029	0.021	0.069
BB%	-0.18	-0.15	-0.12	-0.17	0.22	-0.057	1	-0.0033	0.33	0.21	0.56	-0.085	-0.21	-0.18	0.15	-0.0026	0.055	0.048	0.039	0.017
HR/9	-0.14	-0.13	-0.13	-0.13	0.71	-0.22	-0.0033	1	0.87	0.71	0.35	-0.12	-0.21	-0.17	-0.17	0.28	-0.08	-0.016	-0.036	-0.062
FIP	-0.23	-0.22	-0.21	-0.23	0.83	-0.43	0.33	0.87	1	0.83	0.74	-0.15	-0.29	-0.24	-0.18	0.29	-0.11	-0.0021	-0.019	-0.073
ERA-	-0.2	-0.19	-0.18	-0.19	1	-0.3	0.21	0.71	0.83	1	0.62	-0.13	-0.18	-0.23	-0.094	0.27	-0.048	0.024	0.0057	-0.071
xFIP	-0.27	-0.26	-0.25	-0.27	0.63	-0.58	0.56	0.35	0.74	0.62	1	-0.13	-0.28	-0.24	-0.2	0.18	-0.17	-0.0062	-0.012	-0.06
WPA	0.12	0.068	0.043	0.081	-0.13	0.22	-0.085	-0.12	-0.15	-0.13	-0.13	1	0.48	0.69	0.0023	-0.11	0.089	0.0073	0.019	0.021
WAR	0.73	0.7	0.67	0.72	-0.18	0.35	-0.21	-0.21	-0.29	-0.18	-0.28	0.48	1	0.82	0.092	-0.13	0.16	-0.049	-0.078	0.15
RA9-WAR	0.64	0.59	0.57	0.61	-0.23	0.29	-0.18	-0.17	-0.24	-0.23	-0.24	0.69	0.82	1	0.057	-0.15	0.13	-0.043	-0.063	0.13
Pace (pi)	0.15	0.16	0.17	0.16	-0.093	0.42	0.15	-0.17	-0.18	-0.094	-0.2	0.0023	0.092	0.057	1	-0.11	0.55	-0.0077	-0.027	0.028
HardHit%	-0.12	-0.11	-0.11	-0.11	0.28	-0.13	-0.0026	0.28	0.29	0.27	0.18	-0.11	-0.13	-0.15	-0.11	1	-0.06	0.097	0.075	-0.07
Stuff+	0.14	0.14	0.14	0.14	-0.049	0.44	0.055	-0.08	-0.11	-0.048	-0.17	0.089	0.16	0.13	0.55	-0.06	1	0.061	0.032	0.047
playerid	-0.098	-0.092	-0.092	-0.092	0.025	0.029	0.048	-0.016	-0.0021	0.024	-0.0062	0.0073	-0.049	-0.043	-0.0077	0.097	0.061	1	0.88	-0.12
mlbamid	-0.15	-0.14	-0.14	-0.14	0.0062	0.021	0.039	-0.036	-0.019	0.0057	-0.012	0.019	-0.078	-0.063	-0.027	0.075	0.032	0.88	1	-0.11
Timer Violations	0.2	0.21	0.23	0.21	-0.073	0.069	0.017	-0.062	-0.073	-0.071	-0.06	0.021	0.15	0.13	0.028	-0.07	0.047	-0.12	-0.11	1

Data/Changes/ Missing Variables

For data cleaning of our dataset, we took notice of some variables that needed to be transformed to improve and ensure the quality of our data. For the purpose of modeling and the analysis of our data, we changed all data into numbers. Variables such as "K%", "BB%", and "HardHit%" were converted from percentages to decimals. The innings pitched (IP) variable is recorded as a fraction and needed to be converted to a number. As mentioned above, there are 3 outs in an inning and if a pitcher records only 1 out of the 3 needed to complete an inning pitched then it is recorded as .1; for recording 2 of the necessary 3 outs, a pitcher would be awarded .2 innings pitched. So we converted the IP values ending in 0.1 to 0.333 and IP values ending in 0.2 to 0.666. Recording 0 or 3 outs maintains the IP count to a whole number so no preparation would need to be done to these values. Stuff+ is a statistic measuring pitchers' pitch placements, so it does not apply to position players who have come into the game to pitch. After sorting through the data, the 17 pitchers, who account for 288 observations (or pitches) and are lacking a Stuff+ observation, are all position players meaning they do not regularly take the mound and pitch (*Fig. 8)*. These position players coming into relief accounted for 8 total timer violations. 6 of the 17 position players committed timer violations with David Fry and Ernie Clement each accounting for 2 violations apiece in their single outings as pitchers. In order to deal with the missing Stuff+ variables, we have elected to assign the position players a Stuff+ value of 100 as that is the average amongst all pitchers. Variables that serve as identifiers, such as playerID and mlbamid, were also removed from the summary statistics review. This was done because they were redundant information given pitchers' names were included in the dataset and playerID and mlbamid were not used as part of our analysis.



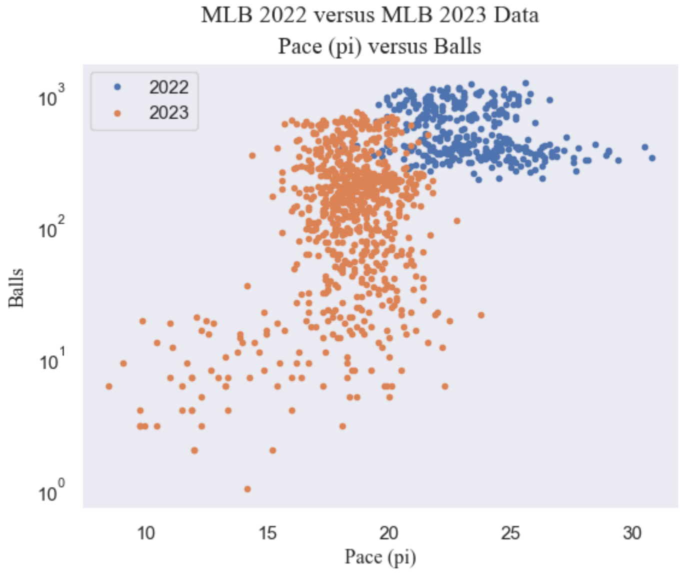
```
In [9]:  df.isnull().sum()

Out[9]: FirstName          0
        LastName           0
        Team               0
        IP                 0
        Pitches            0
        Balls              0
        Strikes            0
        ERA                0
        K%                 0
        BB%                0
        HR/9               0
        FIP                0
        ERA-               0
        xFIP               0
        WPA                0
        WAR                0
        RA9-WAR            0
        Pace (pi)          0
        HardHit%           0
        Stuff+            17
        playerid           0
        mlbamid            0
        TimerViolations    0
        dtype: int64
```

Fig. 8

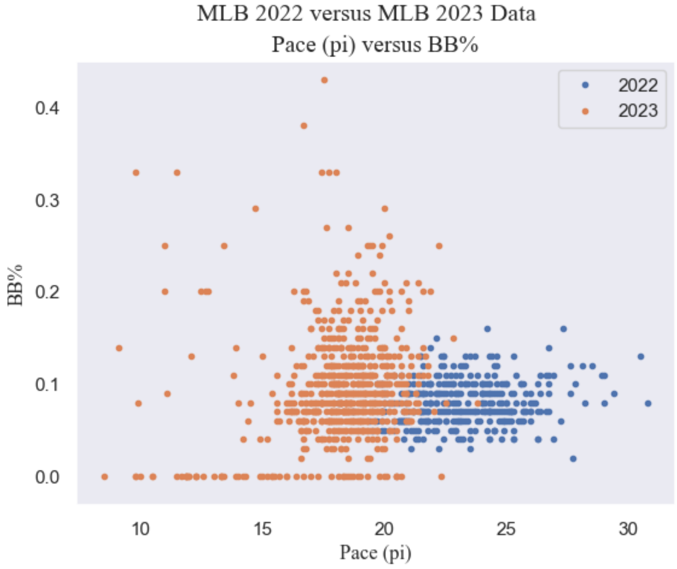
Modeling/Methods

HERE

To start our analysis and modeling of the data, we conducted a K-Nearest Neighbor (KNN) model to further observe the relationship between pitching statistics and Timer Violations. For this analysis, we used the three variables Pitches and ERA- as the independent variables while keeping Timer Violations as the dependent variable. Pitches and ERA- were chosen because these variables contained integer values, because Pitches represents a statistic for the individual pitcher, and because ERA- represents a statistic that compares each pitcher's performance to all other pitchers. Creating a KNN model requires fine-tuning to find the optimal number of clusters, so we utilized the Elbow Method and Silhouette Score Method to help with this process. The Elbow Rule evaluates cluster based on points' distances from their associated cluster center-point, which is referred to as inertia, and compares it to the number of clusters in the dataset. A graph illustrating the Elbow Rule shows at what number of clusters is ideal for a given data. This point is where "line 1" starts to show a more linear relationship, as opposed to quadratic relationship. The graph depicting the Elbow Method for our data can be seen below in *Fig 9*.



MLB 2022 versus MLB 2023 Data
Pace (pi) versus Balls



MLB 2022 versus MLB 2023 Data
Pace (pi) versus BB%

Fig. 9

Because “line 1” becomes more linear around 2 or 3 clusters, we were unsure which number of clusters to use, so we followed with a Silhouette Score Method to determine which number to use. The Silhouette Score measures how well the data has been clustered together on a scale of -1 to 1. A Silhouette Score of -1 indicates poor clustering by means of mismatching; a Silhouette Score of 0 indicates overlapping amongst the clusters; and a Silhouette Score of 1 indicates perfect distribution of the data into clusters. Having a Silhouette Score closer to 1 is ideal. For our dataset, the number of clusters with the highest Silhouette Score was 2, which can be seen in *Fig 10*.

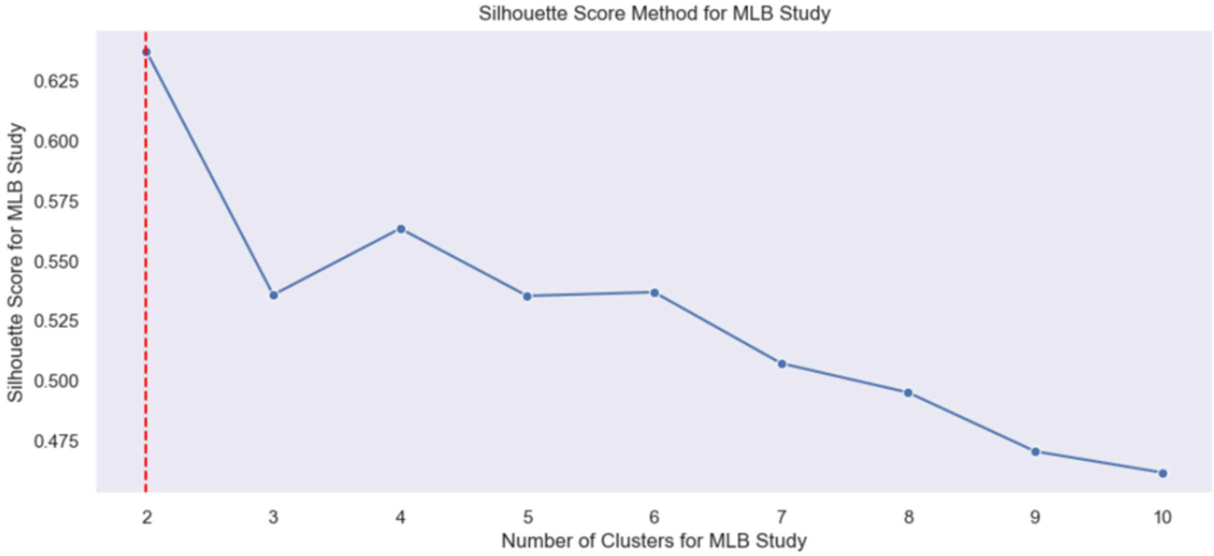


Fig. 10

Knowing that our ideal number of clusters should be 2, we created our KNN model. As we can see in the Pitcher vs. Timer Violations graph, the center of Cluster 0, which consists of all red dots, is around 300 pitches thrown and 0.85 Timer Violations; the center of Cluster 1, which consists of all the blue dots, is around 1375 pitches thrown and 1.10 timer violations. As for the ERA- vs. Timer Violations graph, Cluster 0, once again represented by red dots, has a center around 1400 ERA- and Timer Violations of 1.10; Cluster 1, depicted with blue dots, has a center around 400 ERA- and Timer Violation of 0.85.

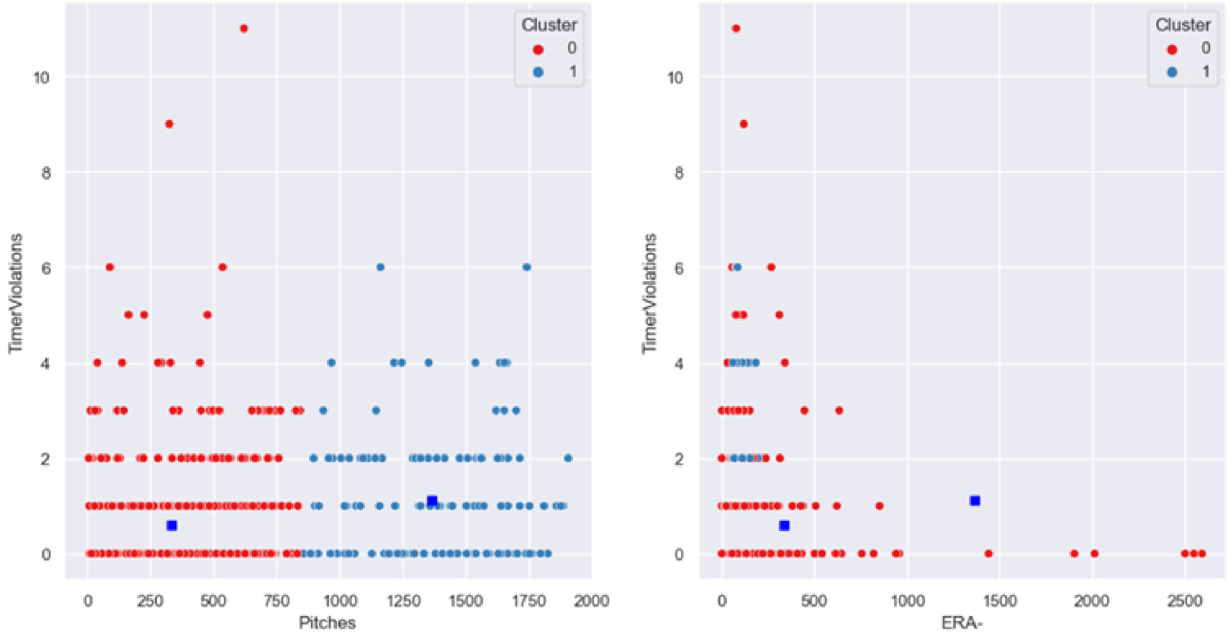
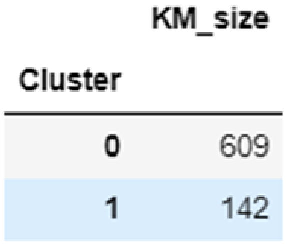


Fig. 11



	KM_size
Cluster	
0	609
1	142

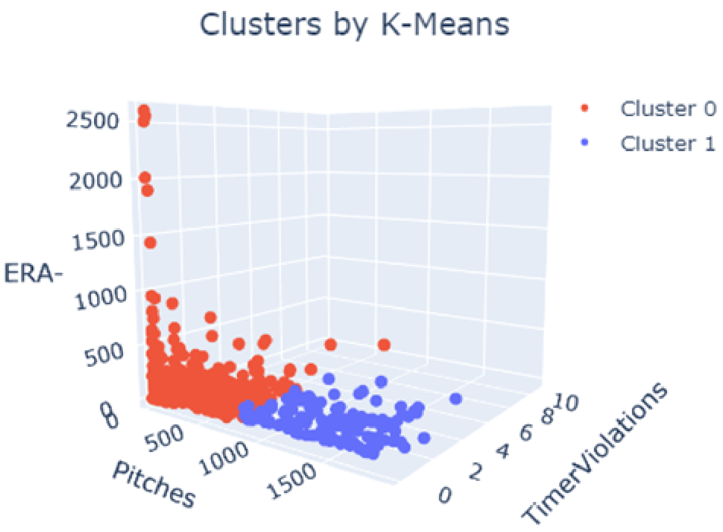


Fig. 12

Looking at the distribution of the datasets within the KNN cluster, we can see that most of the dataset is grouped into Cluster 0 with 609 of 751 combined observations, compared to Cluster 1 with the remaining 142. Additional observations showed the presence of outliers which would explain the skewness of the data. To better understand the effect and relationship of Timer Violations we next conducted a linear regression model.

To predict how many timer violations should have occurred in the 2022 season we created a linear regression model. For this model, we used just the two variables pitches, serving as our independent variable, and timer violations, serving as our dependent variable. In this simple model, we divided our dataset into 66% for the training set and 33% for our testing set. Given the simple linear regression model of $Y=\beta 0+ \beta 1\ X+\varepsilon$, our model produced the following equation: Timer Violations = 0.0004313(Pitches) + 0.47728.

This roughly means that for every pitch, the number of timer violations increases by 0.0004313. With an intercept of 0.477, a pitcher will need to throw 1212 pitches for the model to predict that one timer violation should have occurred.

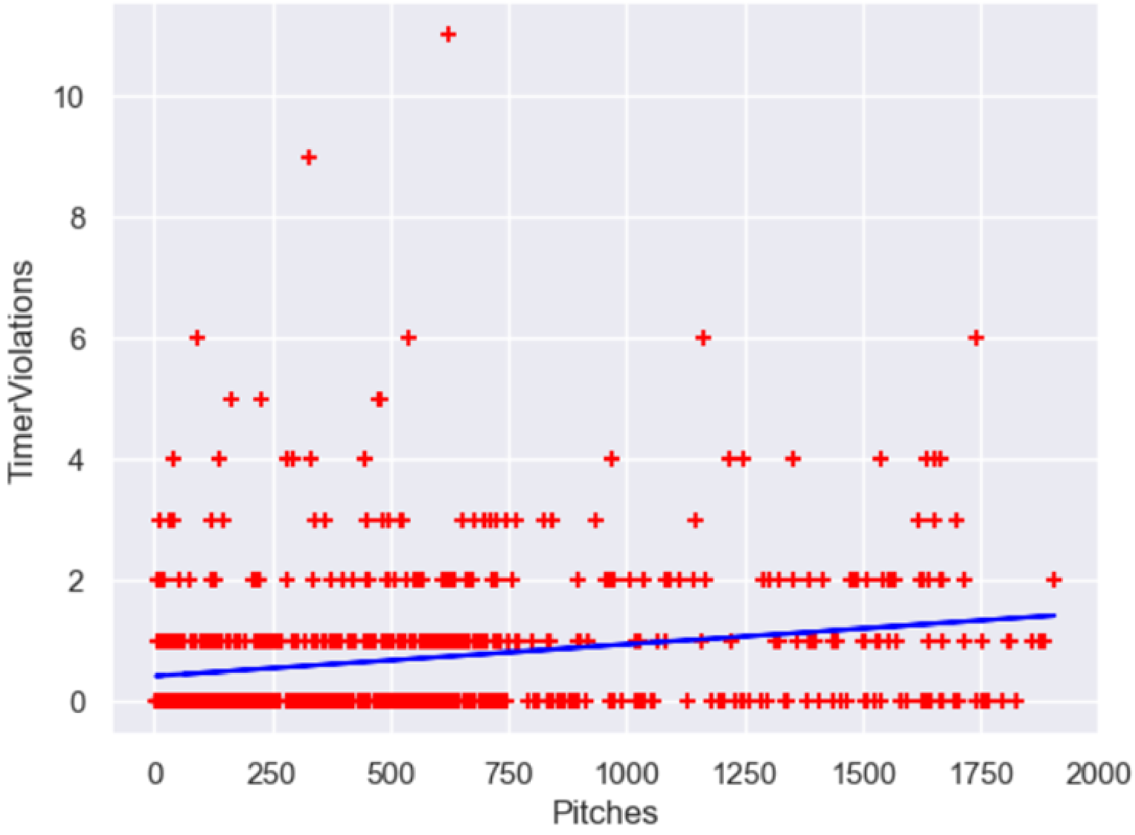


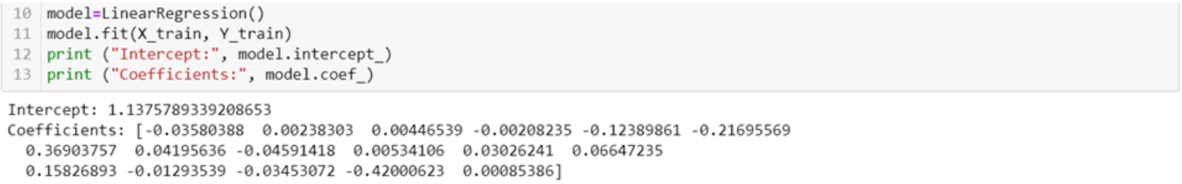
Fig. 13

In the graph above (*Fig. 13)*, each red dot indicates an individual pitcher with the number of pitches he has thrown versus timer violation he has committed. The blue line indicates the linear regression model/equation we created for these two models. As of July 8th, 2023, Kevin Gausman has thrown 1906 pitches; so, using our model Gausman would be predicted to have only thrown 1.299 timer violations. This prediction is lower than the actual number of timer violations Gausman has committed so far in the season with 2. Using this model to predict the number of timer violations for the 2022 season, we would use the 708539 pitches thrown. The results of our model calculated that there should have been 306.072 timer violations over the span of the 2022 season. To determine how well the data fit the model, we calculated the R-Square score of 0.115. Having such a low R-Squared value dictates that the goodness-of-fit between pitches thrown and timer violations is low and not ideal, so there is still high variation among these two variables. The 2022 MLB was unhampered by any game cancellations due to any unforeseen external circumstances like the Covid-19 pandemic or internal circumstances like a players strike, so a total of 4,860 games were played over the course of the regular season. Dividing the total number of games played divided by the predicted Timer Violations in 2022 showed that there should be a Timer Violation for every 15.88 games played. With 30 teams in the MLB, there are typically 15 games played per day which means there should be roughly 1 Timer Violation per day.

Each pitcher has a unique approach to his game, so we built additional linear regression models that would predict the number of timer violations based on all pitcher performance statistics. As mentioned above when introducing and defining the variables, we used the following statistics as independent variables in our modeling: IP, Pitches, Balls, Strikes, ERA, K%, BB%, HR/9, FIP, ERA-, xFIP, WPA, WAR, RA9-WAR, Pace(pi), HardHit%, and Stuff+. The associated dependent variable was timer violations. Like in our first model, we separate the data into a training set and testing set; 66% of the data was used for training while the remaining 33% was used for testing. As a result, our model produced the following equation:

Timer Violations = -0.03580(IP) + 0.00238(Pitches) + 0.00447(Balls) - 0.00208(Strikes) - 0.12390(ERA) - 0.21696(K%) + 0.36904(BB%) + 0.04196(HR/9) – 0.04591(FIP) +

0.00534(ERA-) + 0.03026(xFIP) + 0.06647(WPA) + 0.15827(WAR) – 0.01294(RA9-WAR) – 0.03453(Pace(pi)) – 0.42001(HardHit%) + 0.00085(Stuff+) + 1.13758

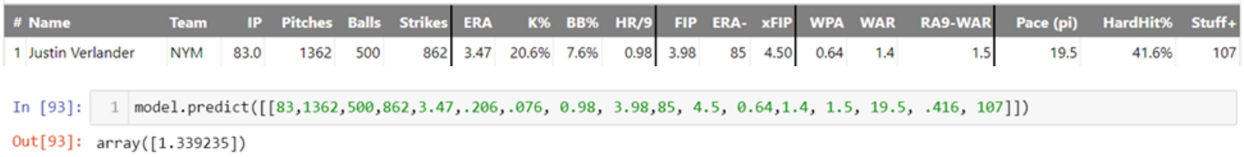


```
model=LinearRegression()
model.fit(X_train, Y_train)
print ("Intercept:", model.intercept_)
print ("Coefficients:", model.coef_)
```

```
Intercept: 1.1375789339208653
Coefficients: [-0.03580388  0.00238303  0.00446539 -0.00208235 -0.12389861 -0.21695569
  0.36903757  0.04195636 -0.04591418  0.00534106  0.03026241  0.06647235
  0.15826893 -0.01293539 -0.03453072 -0.42000623  0.00085386]
```

Fig. 14

With this equation we can input the pitcher statistics and predict the number of timer violations. Looking at pitcher Justin Verlander, who has committed one timer violation, has produced the statistics seen below. Using our model, Justin Verlander should have committed 1.33924 timer violations. Once again in order to gauge the performance of the model, we calculated the R-Squared score of 0.00075, indicating that the variables relating to pitcher performance have next to no relation to the number of timer violations.

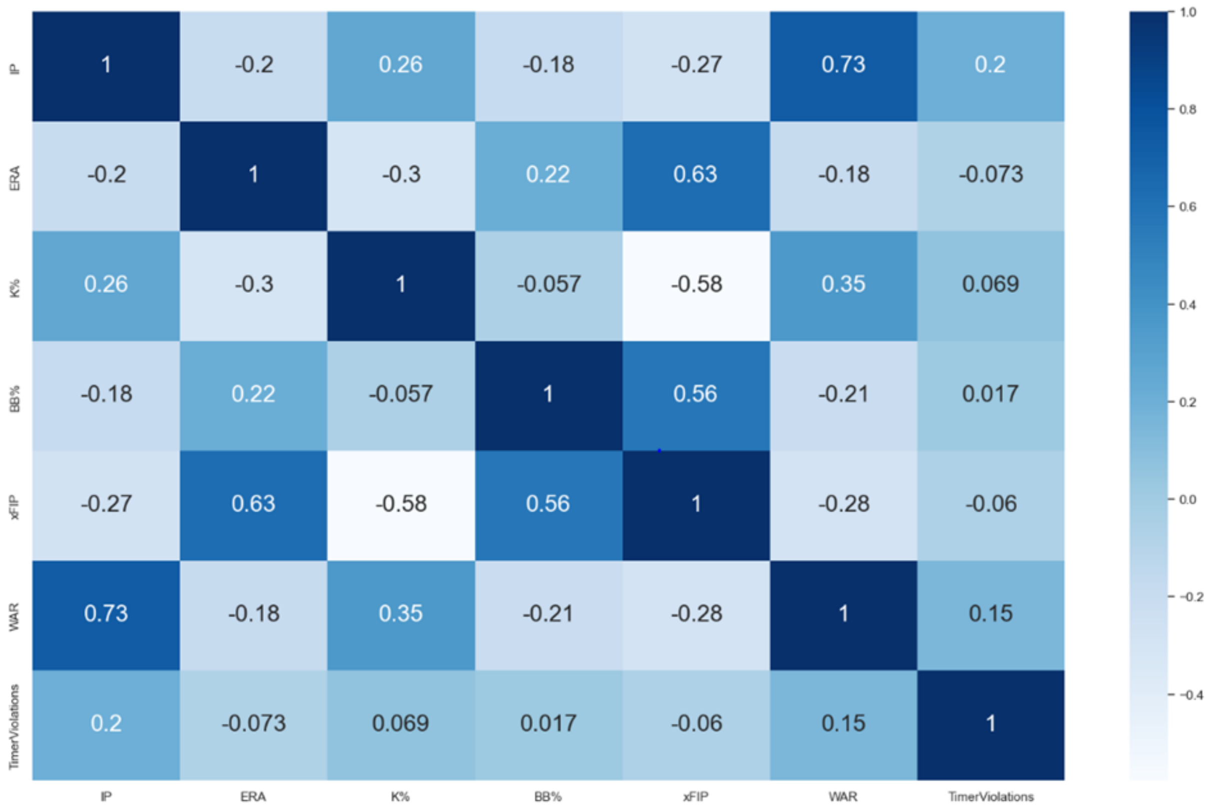


#	Name	Team	IP	Pitches	Balls	Strikes	ERA	K%	BB%	HR/9	FIP	ERA-	xFIP	WPA	WAR	RA9-WAR	Pace (pi)	HardHit%	Stuff+
1	Justin Verlander	NYM	83.0	1362	500	862	3.47	20.6%	7.6%	0.98	3.98	85	4.50	0.64	1.4	1.5	19.5	41.6%	107

```
In [93]:  model.predict([[83,1362,500,862,3.47,.206,.076, 0.98, 3.98,85, 4.5, 0.64,1.4, 1.5, 19.5, .416, 107]])
Out[93]: array([1.339235])
```

Fig. 15

In hopes to create a better linear regression model, we looked to eliminate variables with low correlation and looked at the statistics that would be considered the most important when evaluating pitchers. When evaluating pitchers, many analysts look for IP, ERA, K%, BB%, xFIP, and WAR; so, these were the variables for the next linear regression model. In the graph below, we have provided a heat map with the variable correlations.



	IP	ERA	K%	BB%	xFIP	WAR	TimerViolations
IP	1	-0.2	0.26	-0.18	-0.27	0.73	0.2
ERA	-0.2	1	-0.3	0.22	0.63	-0.18	-0.073
K%	0.26	-0.3	1	-0.057	-0.58	0.35	0.069
BB%	-0.18	0.22	-0.057	1	0.56	-0.21	0.017
xFIP	-0.27	0.63	-0.58	0.56	1	-0.28	-0.06
WAR	0.73	-0.18	0.35	-0.21	-0.28	1	0.15
TimerViolations	0.2	-0.073	0.069	0.017	-0.06	0.15	1

Fig. 16

Using these variables creates the following linear regression equation:

$$\text{TimerViolations} = 0.00583(\text{IP}) - 0.0042014(\text{ERA}) - 0.23207(\text{K\%}) + 1.70934(\text{BB\%}) - 0.02461(\text{xFIP}) + 0.10208(\text{WAR}) + 0.50167$$

To measure the performance of this model we calculated its R-Square score which is 0.04055. This has a higher score than the model using all variables but is still smaller compared to the linear regression model between balls and timer violations.

Evaluation/Discussion

An evaluation of the KNN and linear regression models of the data shows us there is not statistically significant correlation between pitch timer violations and pitcher performance. The most significant correlation between timer violations and any other variable was the total number of balls thrown by the pitcher. This makes complete sense, as according to the rule, any timer violation from the pitcher results in an automatic ball without a pitch being thrown.

Even for pitchers who are extreme outliers with their total timer violations, there is not a statistically significant drop in performance. Craig Kimbrel is the pitcher with the most timer violations of 2023, with 11. The next closest is Anthony Bass, with 9 violations. Let's take these two pitchers and compare their performances in 2023 compared to 2022. In 63 innings in 2022, Kimbrel had a 3.75 ERA with a 10.8 Walk%, a 27.7% Strikeout%, and 0.6 home runs/9. So far in 2023, through 40 innings, he has a 3.15 ERA with a 10.2 Walk%, 38.2 strikeout%, and 1.13 home runs/9. So what we can take from this is that Kimbrel has been better at preventing runs and striking out hitters, despite allowing more home runs, and his walk rate has remained steady to his 2022 performance. The large presence of outliers like Kimbrel means that the variables used to measure pitcher performance have proven to not be related to timer violations. "Intangibles" is a term that is used in sports regarding something that cannot be taught or measured such as attitude, talent, effort, etc.; although pitching mechanics can be taught and refined, the approach to an at-bat is highly unique, so "Timer Violations" might be considered an intangible statistic for pitchers.

Our results showing no statistically significant correlation between timer violations and pitcher performance is actually great news for Major League Baseball. Going into the season, there was some concern among both fans and media about what impact the pitch clock would have. Many thought it might make pitchers less effective as they might feel rushed and have less time to focus, get in a rhythm, and decide on what pitch to throw. Sports fans, especially baseball fans, are very much traditionalists and hate change. But overall, MLB’s goals have been accomplished with the new rules. Games are faster, stolen bases are up, and MLB is seeing higher viewership and attendance than they have in previous years. MLB had tested the timer for years in the minor leagues before it came to the major leagues, but fans who do not follow the minor leagues were seeing it for the first time. There was plenty of hysteria in the first week of spring training after there were a few pitch timer violations in high leverage situations. Nobody wants to see a game end because of a timer violation, and fans were concerned about the playoffs and World Series being decided because of the pitch timer. In July, the MLB Players Association expressed an interest in tweaking the timer rules just for the playoffs, but it was unclear if MLB would take action.

There are still plenty of opportunities for future study on this topic. One such study could be conducted on the impact the pitch clock has had on pitcher injuries. In May of 2023, Travis Sawchik of The Score explored the increase in pitcher elbow injuries from 2022 to 2023. His study found that the number of pitchers going on the Injured List was higher than the previous season, although the number of total days spent on the injured list was slightly lower. Much more information over a longer period of time to see if there is any significant correlation between injuries and a pitch timer. Injuries are random and hard to predict, so an in depth study would be required before we can attribute a pitch timer to the cause of an increase of injuries.

Another opportunity for future research is if the number of pitch timer violations has decreased as the season has progressed. Our data did not include spring training data, nor did it include the date the violation occurred. Much of the hysteria around the pitch clock was in the first week of spring training. The pitch clock was new to veteran pitchers who had not played under the rules in the minor leagues because they had already reached the pros by the time the rules were implemented for testing. After the first week, the online discussion around the timer seemed to lessen. It is likely pitchers needed a few games to get used to the new rules, and it makes logical sense for this to be the case.

On average, when accounting for a full season, our calculations estimated an average of one violation per day over the 162 game season. With up to 15 games per day, the effect of the pitch clock timer is basically non-existent on pitcher performance. Ultimately, MLB has been extremely successful with this rule change. Games are 30 minutes faster, pitcher performance is the same as it's always been, stolen bases are up, and most importantly, viewership is up as well.

Conclusion

Like everything else in the world, the game of baseball is ever changing. With the inception of the most recent season, the MLB brought sweeping changes, one of which came in the form of the pitch clock. Throughout our analysis of data from last season and this season, we determined that the implementation of the pitch clock and the penalty for taking too much time, which we measured by the variable of Timer Violation, has no correlation or effect to the performance of the pitchers. Each violation is too sporadic or situational to discern a concrete pattern/conclusion as to which variables may lead to them. The lack of an effect would be deemed a positive to MLB owners whose goal has been to shorten the length of games. As of July 8th, 2023, no game has been decided due to a Timer Violation.

Works Cited

Associated Press. (2023, March 30). MLB rule changes: pitch clock, larger bases and more. Retrieved from Associated Press: https://apnews.com/article/mlb-rule-changes-explainer-61dc5754b7e14da5bf62074a26e3c07b

BETMGM. (2023, April 03). How Long Do Baseball Games Last? Retrieved from The Roar: https://sports.betmgm.com/en/blog/mlb/how-long-do-baseball-games-last-bm06/

Earned Run (ER): Glossary. MLB.com. (n.d.). https://www.mlb.com/glossary/standard-stats/earned-run

Hook, C. (2022, August 18). An introduction to seam-shifted wakes and their effect on sinkers. Driveline Baseball. https://www.drivelinebaseball.com/2020/11/more-than-what-it-seams-an-introduction-to-seam-shifted-wakes-and-their-effect-on-sinkers/

McGrattan, O. (2023, March 10). Stuff+, location+, and pitching+ primer. Sabermetrics Library. https://library.fangraphs.com/pitching/stuff-location-and-pitching-primer/

Merriam-Webster. (2023). Baseball. Retrieved from Merriam-Webster: https://www.merriam-webster.com/dictionary/baseball#:~:text=base%C2%B7%E2%80%

8Bball%20%CB%88b%C4%81s%2D%CB%8Cb%C8%AFl,ball%20used%20in%20this%20game

MLB. (2023). Strike Zone Definition. Retrieved from MLB.com: https://www.mlb.com/glossary/rules/strike-zone

Rubin, M. (2023, May). How long is a baseball game in 2023? Effects of new MLB rule changes. Retrieved from Fansided: https://fansided.com/2023/05/23/how-mlb-games-2023-effects-mlb-rule-changes/#:~:text=How%20long%20are%20MLB%20games,three%20hours%20and%20three%20minutes

Rules of Sports. (2022). Baseball Rules. Retrieved from Rules of Sports: https://www.rulesofsport.com/sports/baseball.html

Seattle University. (n.d.). Physics. Seattle University. https://www.seattleu.edu/scieng/physics/physics-demos/thermodynamics/magnus-effect/

Slowinski, P. (2010, February 15). FIP. Sabermetrics Library. https://library.fangraphs.com/pitching/fip/

Slowinski, P. (2010b, February 15). XFIP. Sabermetrics Library. https://library.fangraphs.com/pitching/xfip/

Slowinski, P. (2010c, February 16). WPA. Sabermetrics Library. https://library.fangraphs.com/misc/wpa/

Slowinski, P. (2011, April 8). Era- / FIP- / xfip-. Sabermetrics Library. https://library.fangraphs.com/pitching/era-fip-xfip/

Slowinski, P. (2012, March 16). War. Sabermetrics Library. https://library.fangraphs.com/war/

Snyder, M. (2023, March 30). MLB pitch clock: New baseball rule explained, how it worked in spring training and who could be affected. Retrieved from CBS Sports: https://www.cbssports.com/mlb/news/mlb-pitch-clock-new-baseball-rule-explained-how-it-worked-in-spring-training-and-who-could-be-affected/

Weinberg, N. (2015, May 10). Quality of Contact stats. Sabermetrics Library. https://library.fangraphs.com/pitching/quality-of-contact-stats/