```
import pandas as pd
import numpy as np
from pandas.plotting import scatter_matrix
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.model_selection import KFold
from sklearn.model_selection import cross_val_score
import seaborn as sns
sns.set(color_codes=True)
import matplotlib.pyplot as plt
import os
os.chdir('C:\\Users\\Mauricio Sanchez\\Downloads')
os.getcwd()
```

```
'C:\\Users\\Mauricio Sanchez\\Downloads'
```

```
# Specify the location of the dataset.
MLB = 'PitchingVolations.csv'


# Load the data into a Pandas DataFrame
df= pd.read_csv (MLB, header=None)


df.head()
```

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | ... | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | Jose | A. Ferrer | WSN | 4.1 | 63 | 29 | 34 | 0.00 | 0.31 | 0.19 | ... | 3.43 | 0.07 | 0.1 | 0.3 | 18.9 | 0.50 | 72.0 | 24017 | 678606 |
| **1** | Fernando | Abad | COL | 5.0 | 116 | 46 | 70 | 5.40 | 0.08 | 0.08 | ... | 7.20 | -0.02 | -0.1 | 0.0 | 19.2 | 0.41 | 57.0 | 4994 | 472551 |
| **2** | Andrew | Abbott | CIN | 41.2 | 674 | 240 | 434 | 2.38 | 0.29 | 0.08 | ... | 4.37 | 1.20 | 0.9 | 1.8 | 17.9 | 0.43 | 87.0 | 29911 | 671096 |
| **3** | Cory | Abbott | WSN | 17.0 | 294 | 112 | 182 | 4.24 | 0.18 | 0.12 | ... | 4.85 | -0.25 | 0.0 | 0.1 | 18.1 | 0.35 | 97.0 | 20277 | 676265 |
| **4** | Bryan | Abreu | HOU | 42.0 | 746 | 282 | 464 | 2.79 | 0.37 | 0.10 | ... | 2.89 | -0.09 | 0.6 | 0.8 | 19.9 | 0.34 | 139.0 | 16609 | 650556 |

```
col_names = ['FirstName', 'LastName', 'Team', 'IP', 'Pitches','Balls', 'Strikes', 'ERA', 'K%', 'BB%', 'HR/9', 'FIP', 'ERA-', 'xFIP', 'WPA' , 'WAR', 'RA9-WAR', 'Pace (pi)', 'HardHit%', 'St


df.columns = col_names


# Look at the first 5 rows of data
df.head()
```

| | FirstName | LastName | Team | IP | Pitches | Balls | Strikes | ERA | K% | BB% | ... | xFIP | WPA | WAR | RA9-WAR | Pace (pi) | HardHit% |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Jose | A. Ferrer | WSN | 4.1 | 63 | 29 | 34 | 0.00 | 0.31 | 0.19 | | 3.43 | 0.07 | 0.1 | 0.3 | 18.9 | 0.50 |

```
df.isnull().sum()
```

```
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
Timer Violations     0
dtype: int64
```

```
print(df.shape)
```

```
(751, 23)
```

```
print(df.dtypes)
```

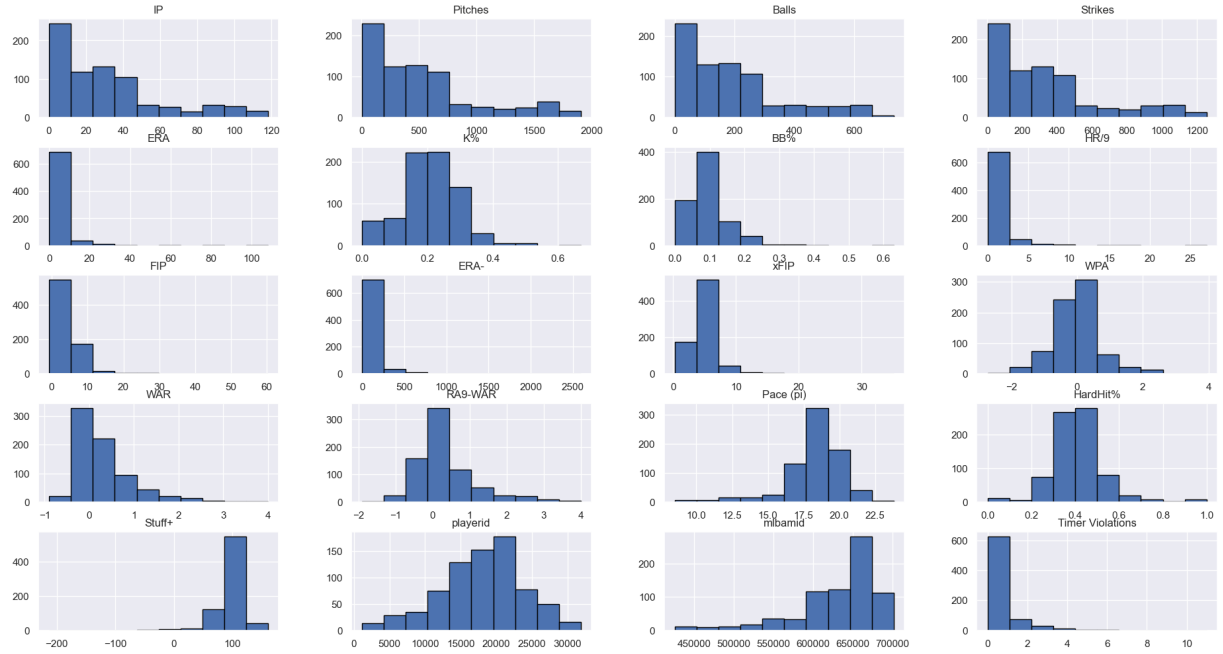
```
FirstName           object
LastName            object
Team                object
IP                 float64
Pitches              int64
Balls                int64
Strikes              int64
ERA                float64
K%                 float64
BB%                float64
HR/9               float64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

```
Timer Violations      int64
dtype: object
```

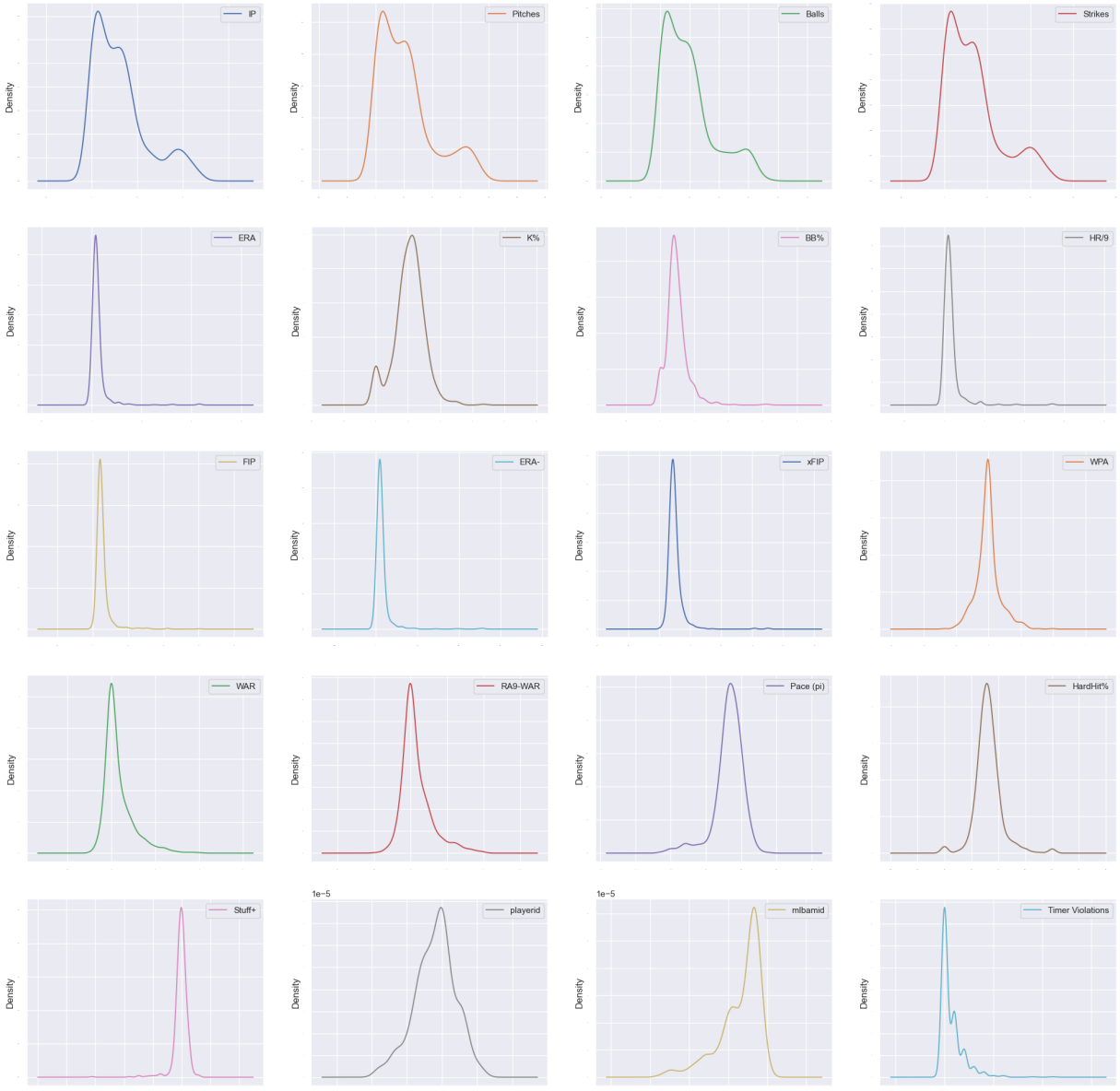
```
print(df.describe())
```

```
                IP      Pitches       Balls      Strikes         ERA  \
count  751.000000   751.000000  751.000000   751.000000  751.000000
mean    31.479893   528.605859  191.234354   337.371505    5.997909
std     29.393286   478.460422  171.259673   308.717164    9.177736
min      0.100000     4.000000    0.000000     3.000000    0.000000
25%      7.150000   137.500000   52.000000    83.500000    3.035000
50%     25.000000   423.000000  154.000000   273.000000    4.310000
75%     42.100000   697.000000  253.000000   445.500000    6.170000
max    118.100000  1906.000000  733.000000  1255.000000  108.000000

                K%         BB%        HR/9         FIP         ERA-  \
count  751.000000  751.000000  751.000000  751.000000   751.000000
mean     0.207044    0.095792    1.484927    5.242836   140.025300
std      0.092906    0.063023    2.389082    4.389534   216.792443
min      0.000000    0.000000    0.000000   -0.710000     0.000000
25%      0.160000    0.060000    0.520000    3.410000    70.000000
50%      0.210000    0.090000    1.060000    4.360000   101.000000
75%      0.260000    0.120000    1.660000    5.515000   141.000000
max      0.670000    0.630000   27.000000   60.290000  2593.000000

              xFIP         WPA         WAR     RA9-WAR   Pace (pi)    HardHit%  \
count  751.000000  751.000000  751.000000  751.000000  751.000000  751.000000
mean     4.974434   -0.034394    0.316112    0.315712   18.245273    0.406884
std      2.685985    0.725944    0.659712    0.830096    2.073123    0.122728
min      0.290000   -2.720000   -0.900000   -1.900000    8.500000    0.000000
25%      3.825000   -0.370000   -0.100000   -0.100000   17.500000    0.350000
50%      4.540000   -0.020000    0.100000    0.100000   18.500000    0.400000
75%      5.355000    0.175000    0.500000    0.600000   19.500000    0.460000
max     35.080000    3.940000    4.000000    4.000000   23.800000    1.000000

            Stuff+      playerid        mlbamid  Timer Violations
count  734.000000    751.000000     751.000000        751.000000
mean    96.694823  17989.014647  630512.246338          0.699068
std     25.373705   5778.541275   53869.292443          1.168469
min   -213.000000   1157.000000  425794.000000          0.000000
25%     90.000000  14462.500000  605473.000000          0.000000
50%     99.000000  18454.000000  656266.000000          0.000000
75%    108.000000  21508.500000  669003.000000          1.000000
max    161.000000  31839.000000  701643.000000         11.000000
```

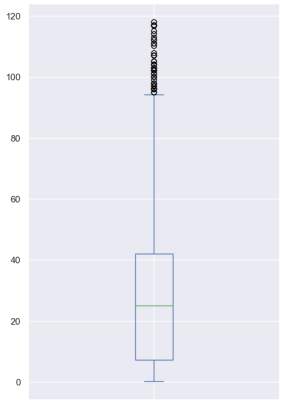
```
df.hist(edgecolor= 'black',figsize=(23,12))
plt.show()
```

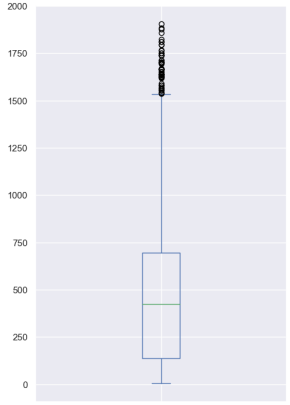


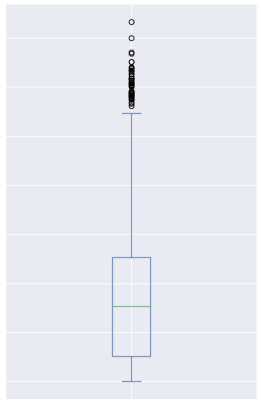

```
df.plot(kind='density', subplots=True, layout= (5,4), sharex=False,legend=True, fontsize=1, figsize= (25,25))
plt.show()
```

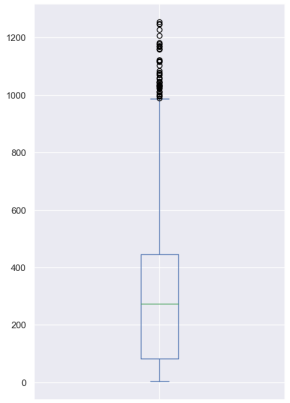

IP
Pitches
Balls
Strikes
ERA
K%
BB%
HR/9
FIP
ERA-
xFIP
WPA
WAR
RA9-WAR
Pace (pi)
HardHit%
Stuff+
playerid
mlbamid
Timer Violations
Density
1e−5

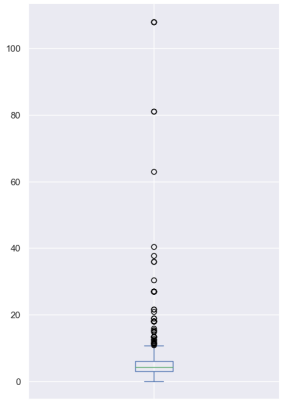
```
df.plot(kind="box", subplots=True, layout=(5,4), sharex=False, figsize=(25,50))
plt.show()
```

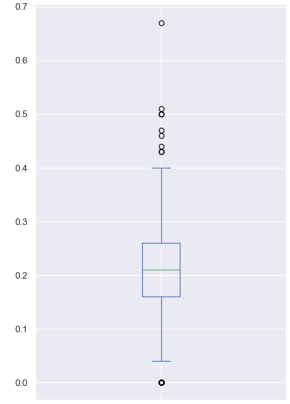

IP

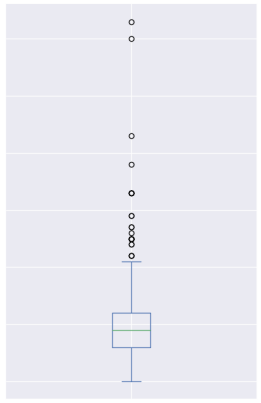

Pitches

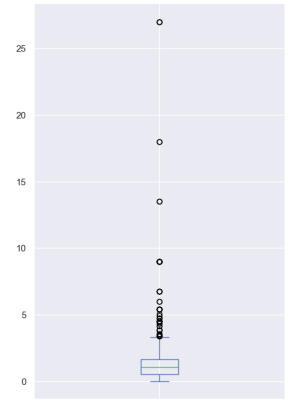

Balls

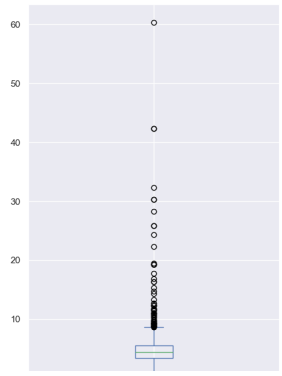

Strikes

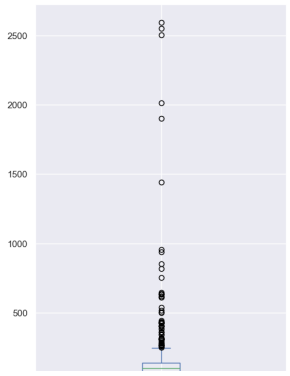

ERA

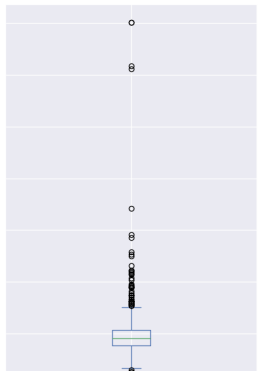

K%

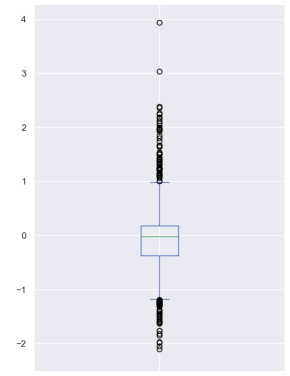

BB%


HR/9









```
pd.options.display.float_format = '{:,.3f}'.format
```

```
df.corr()
```

```
C:\Users\LordG\AppData\Local\Temp\ipykernel_22804\1134722465.py:1: FutureWarning: The default value of numeric_only in
  df.corr()
```

| | IP | Pitches | Balls | Strikes | ERA | K% | BB% | HR/9 | FIP | ERA- | xFIP | WPA | WAR | RA9-WAR | Pace (pi) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **IP** | 1.000 | 0.995 | 0.982 | 0.996 | -0.200 | 0.256 | -0.181 | -0.139 | -0.232 | -0.198 | -0.272 | 0.117 | 0.730 | 0.640 | 0.148 |
| **Pitches** | 0.995 | 1.000 | 0.994 | 0.998 | -0.189 | 0.261 | -0.153 | -0.132 | -0.222 | -0.187 | -0.262 | 0.068 | 0.701 | 0.594 | 0.161 |
| **Balls** | 0.982 | 0.994 | 1.000 | 0.986 | -0.183 | 0.253 | -0.119 | -0.130 | -0.212 | -0.182 | -0.248 | 0.043 | 0.666 | 0.565 | 0.168 |
| **Strikes** | 0.996 | 0.998 | 0.986 | 1.000 | -0.191 | 0.264 | -0.172 | -0.132 | -0.227 | -0.189 | -0.269 | 0.081 | 0.717 | 0.607 | 0.156 |
| **ERA** | -0.200 | -0.189 | -0.183 | -0.191 | 1.000 | -0.302 | 0.218 | 0.707 | 0.829 | 0.999 | 0.629 | -0.127 | -0.182 | -0.228 | -0.093 |
| **K%** | 0.256 | 0.261 | 0.253 | 0.264 | -0.302 | 1.000 | -0.057 | -0.224 | -0.428 | -0.298 | -0.576 | 0.225 | 0.354 | 0.287 | 0.420 |
| **BB%** | -0.181 | -0.153 | -0.119 | -0.172 | 0.218 | -0.057 | 1.000 | -0.003 | 0.328 | 0.214 | 0.563 | -0.085 | -0.208 | -0.175 | 0.153 |

```
plt.figure(figsize =(16,10))
sns.heatmap(df.corr(), annot=True)
plt.show()
```

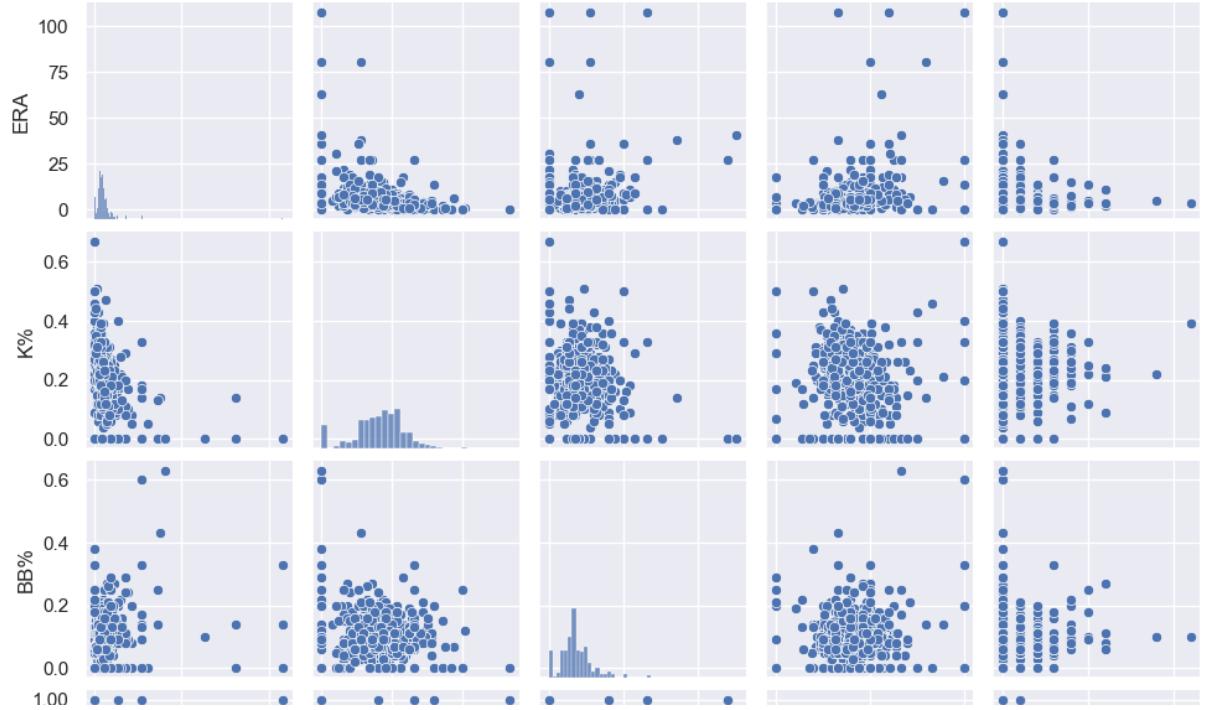
```
C:\Users\LordG\AppData\Local\Temp\ipykernel_22804\1236886484.py:2: FutureWarning: The default value of numeric_only in
  sns.heatmap(df.corr(), annot=True)
```


```
df2= df[['ERA','K%', 'BB%', 'HardHit%','Timer Violations']]
```

```
df2.corr()
```

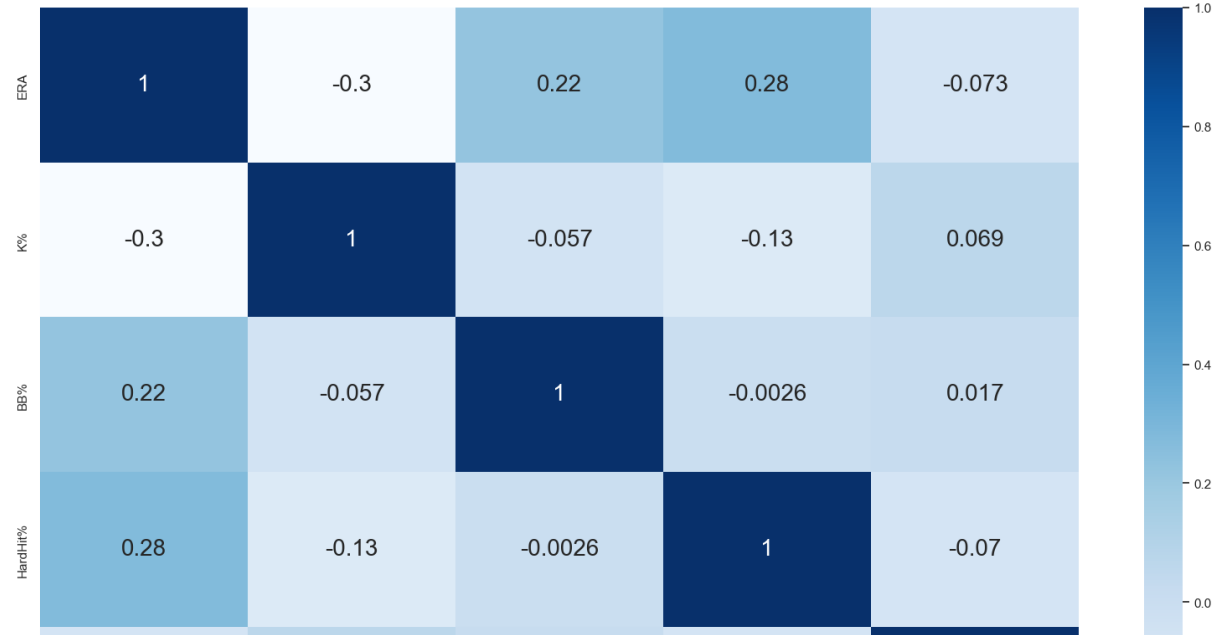
| | ERA | K% | BB% | HardHit% | Timer Violations |
|---|---|---|---|---|---|
| **ERA** | 1.000 | -0.302 | 0.218 | 0.277 | -0.073 |
| **K%** | -0.302 | 1.000 | -0.057 | -0.127 | 0.069 |
| **BB%** | 0.218 | -0.057 | 1.000 | -0.003 | 0.017 |
| **HardHit%** | 0.277 | -0.127 | -0.003 | 1.000 | -0.070 |
| **Timer Violations** | -0.073 | 0.069 | 0.017 | -0.070 | 1.000 |

```
sns.pairplot(df2, height=2);
plt.show()
```




```
plt.figure(figsize =(20,12))
sns.heatmap(df2.corr(), annot=True)
plt.show()
```


```
plt.figure(figsize =(20,12))
sns.heatmap(df2.corr(), cmap="Blues", annot=True, annot_kws={"fontsize":20})
plt.show()
```




```
df2 = df[['Pace (pi)','K%']]
```

```
df3 = df[['ERA','xFIP']]
```

```
df2.head()
```

| | Pace (pi) | K% |
|---|---|---|
| **0** | 18.9 | 0.31 |
| **1** | 19.2 | 0.08 |
| **2** | 17.9 | 0.29 |
| **3** | 18.1 | 0.18 |
| **4** | 19.9 | 0.37 |

```
df3.head()
```

| | ERA | xFIP |
|---|---|---|
| **0** | 0.00 | 3.43 |
| **1** | 5.40 | 7.20 |
| **2** | 2.38 | 4.37 |
| **3** | 4.24 | 4.85 |
| **4** | 2.79 | 2.89 |