nommesen_april-week1-eda

November 9, 2022

[28]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

1 Load the datasets downloaded from Kaggle

[2]:
```
df_2022_train = pd.read_csv('../data/raw/2022_train.csv')
df_2022_test = pd.read_csv('../data/raw/2022_test.csv')
```

[3]:
```
df_2022_train.head()
```

[3]:

	Id	GP	MIN	PTS	FGM	FGA	FG%	3P Made	3PA	3P%	…	FTA	FT%	\
0	3799	80	24.3	7.8	3.0	6.4	45.7	0.1	0.3	22.6	…	2.9	72.1	
1	3800	75	21.8	10.5	4.2	7.9	55.1	-0.3	-1.0	34.9	…	3.6	67.8	
2	3801	85	19.1	4.5	1.9	4.5	42.8	0.4	1.2	34.3	…	0.6	75.7	
3	3802	63	19.1	8.2	3.5	6.7	52.5	0.3	0.8	23.7	…	1.5	66.9	
4	3803	63	17.8	3.7	1.7	3.4	50.8	0.5	1.4	13.7	…	0.5	54.0	

	OREB	DREB	REB	AST	STL	BLK	TOV	TARGET_5Yrs
0	2.2	2.0	3.8	3.2	1.1	0.2	1.6	1
1	3.6	3.7	6.6	0.7	0.5	0.6	1.4	1
2	0.6	1.8	2.4	0.8	0.4	0.2	0.6	1
3	0.8	2.0	3.0	1.8	0.4	0.1	1.9	1
4	2.4	2.7	4.9	0.4	0.4	0.6	0.7	1

[5 rows x 21 columns]

[4]:
```
df_2022_test.head()
```

[4]:

	Id	GP	MIN	PTS	FGM	FGA	FG%	3P Made	3PA	3P%	FTM	FTA	FT%	\
0	0	56	9.1	4.0	1.6	3.7	43.7	0.1	0.3	7.3	0.7	1.2	63.4	
1	1	43	19.3	10.1	3.7	8.1	46.0	0.6	1.7	35.1	1.8	2.5	75.3	
2	2	82	33.9	11.3	4.9	10.6	45.6	0.5	1.9	44.8	1.8	2.7	71.2	
3	3	86	44.7	18.8	6.8	15.9	42.9	0.5	1.8	13.5	4.5	6.3	70.9	
4	4	58	12.3	4.7	1.6	4.0	40.0	0.5	1.7	38.7	1.1	1.3	76.9	

	OREB	DREB	REB	AST	STL	BLK	TOV
0	1.2	0.8	1.7	0.4	0.2	0.3	0.8
1	0.5	0.9	1.5	3.5	0.6	0.0	1.8
2	1.3	3.3	4.5	2.5	1.3	0.3	2.0
3	1.5	3.2	5.0	4.1	0.9	0.1	3.6
4	0.2	0.6	0.9	1.5	0.5	-0.4	0.9

2 1. Structure investigation

The raw train datasets have 8000 rows and 21 columns. The test data set have 3799 rows and 20 columns. The train data set have the target variable. They both have all columns numerical type.

[5]:
```
df_2022_train.shape
```

[5]:
```
(8000, 21)
```

[6]:
```
df_2022_test.shape
```

[6]:
```
(3799, 20)
```

[7]:
```
df_2022_train.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8000 entries, 0 to 7999
Data columns (total 21 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Id           8000 non-null   int64
 1   GP           8000 non-null   int64
 2   MIN          8000 non-null   float64
 3   PTS          8000 non-null   float64
 4   FGM          8000 non-null   float64
 5   FGA          8000 non-null   float64
 6   FG%          8000 non-null   float64
 7   3P Made      8000 non-null   float64
 8   3PA          8000 non-null   float64
 9   3P%          8000 non-null   float64
 10  FTM          8000 non-null   float64
 11  FTA          8000 non-null   float64
 12  FT%          8000 non-null   float64
 13  OREB         8000 non-null   float64
 14  DREB         8000 non-null   float64
 15  REB          8000 non-null   float64
 16  AST          8000 non-null   float64
 17  STL          8000 non-null   float64
```

```
 18  BLK          8000 non-null   float64
 19  TOV          8000 non-null   float64
 20  TARGET_5Yrs  8000 non-null   int64
dtypes: float64(18), int64(3)
memory usage: 1.3 MB
```

[8]:
```
df_2022_test.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3799 entries, 0 to 3798
Data columns (total 20 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   Id       3799 non-null   int64
 1   GP       3799 non-null   int64
 2   MIN      3799 non-null   float64
 3   PTS      3799 non-null   float64
 4   FGM      3799 non-null   float64
 5   FGA      3799 non-null   float64
 6   FG%      3799 non-null   float64
 7   3P Made  3799 non-null   float64
 8   3PA      3799 non-null   float64
 9   3P%      3799 non-null   float64
 10  FTM      3799 non-null   float64
 11  FTA      3799 non-null   float64
 12  FT%      3799 non-null   float64
 13  OREB     3799 non-null   float64
 14  DREB     3799 non-null   float64
 15  REB      3799 non-null   float64
 16  AST      3799 non-null   float64
 17  STL      3799 non-null   float64
 18  BLK      3799 non-null   float64
 19  TOV      3799 non-null   float64
dtypes: float64(18), int64(2)
memory usage: 593.7 KB
```

Create a copy of each of the data to preserve the original data. Rename the features into something more descriptive!

[9]:
```
df_main = df_2022_train.rename(columns = {
    'GP':'Games Played',
    'MIN':'Minutes Played',
    'PTS':'Points Per Game',
    'FGM' : 'Field Goals Made',
    'FGA' : 'Field Goals Attempts',
    'FG%' : 'Field Goals Percent',
    '3P Made' : '3Points Made',
    '3PA' : '3Points Attempts',
```

```
    '3P%' : '3Points Percent',
    'FTM' : 'Free Throw Made',
    'FTA' : 'Free Throw Attempts',
    'FT%' : 'Free Throw Percent',
    'OREB' : 'Offensive Rebounds',
    'DREB' : 'Defensive Rebounds',
    'REB' : 'Rebounds',
    'AST' : 'Assists',
    'STL' : 'Steals',
    'BLK' : 'Blocks',
    'TOV' : 'Turnovers'
    })

df_main.head(20)
```

[9]:

	Id	Games Played	Minutes Played	Points Per Game	Field Goals Made	\
0	3799	80	24.3	7.8	3.0	
1	3800	75	21.8	10.5	4.2	
2	3801	85	19.1	4.5	1.9	
3	3802	63	19.1	8.2	3.5	
4	3803	63	17.8	3.7	1.7	
5	3804	88	20.0	8.8	3.7	
6	3805	70	20.6	7.0	3.2	
7	3806	57	17.6	5.4	1.7	
8	3807	46	26.5	9.3	3.1	
9	3808	64	33.4	16.9	6.4	
10	3809	72	30.6	13.2	4.9	
11	3810	47	15.2	6.5	2.5	
12	3811	41	9.8	2.3	0.9	
13	3812	64	29.7	12.0	4.2	
14	3813	90	31.1	12.9	4.9	
15	3814	45	5.2	1.3	0.5	
16	3815	29	7.5	1.9	0.7	
17	3816	53	13.8	6.7	2.1	
18	3817	83	8.5	2.2	0.9	
19	3818	64	10.4	5.1	2.2	

	Field Goals Attempts	Field Goals Percent	3Points Made	3Points Attempts	\
0	6.4	45.7	0.1	0.3	
1	7.9	55.1	-0.3	-1.0	
2	4.5	42.8	0.4	1.2	
3	6.7	52.5	0.3	0.8	
4	3.4	50.8	0.5	1.4	
5	8.7	43.6	-0.2	-0.5	
6	5.4	58.9	0.1	-0.1	
7	4.7	37.2	0.9	2.8	
8	7.2	44.0	0.6	1.7	

9	11.3	59.0	0.3	1.1
10	11.0	46.9	1.0	2.5
11	6.5	38.3	0.7	2.5
12	2.3	37.9	-0.4	-0.6
13	9.6	45.2	0.8	2.1
14	10.3	47.0	0.3	1.1
15	1.4	41.0	0.1	0.1
16	2.6	30.1	0.3	1.0
17	4.7	43.5	0.2	0.7
18	1.8	48.4	-0.3	-1.0
19	4.9	45.5	-0.1	-0.2

	3Points Percent	…	Free Throw Attempts	Free Throw Percent	\
0	22.6	…	2.9	72.1	
1	34.9	…	3.6	67.8	
2	34.3	…	0.6	75.7	
3	23.7	…	1.5	66.9	
4	13.7	…	0.5	54.0	
5	6.9	…	2.8	75.1	
6	-3.6	…	1.4	56.7	
7	33.2	…	1.8	78.3	
8	27.5	…	2.3	82.1	
9	-13.0	…	4.9	74.9	
10	48.1	…	2.8	74.1	
11	39.7	…	1.1	75.5	
12	-4.7	…	1.3	64.7	
13	28.3	…	3.3	76.8	
14	17.0	…	2.6	93.4	
15	21.1	…	0.2	73.3	
16	26.8	…	0.4	68.0	
17	19.8	…	3.8	75.1	
18	5.6	…	0.9	63.4	
19	-6.1	…	1.8	43.6	

	Offensive Rebounds	Defensive Rebounds	Rebounds	Assists	Steals	Blocks	\
0	2.2	2.0	3.8	3.2	1.1	0.2	
1	3.6	3.7	6.6	0.7	0.5	0.6	
2	0.6	1.8	2.4	0.8	0.4	0.2	
3	0.8	2.0	3.0	1.8	0.4	0.1	
4	2.4	2.7	4.9	0.4	0.4	0.6	
5	1.8	2.9	4.7	1.8	0.4	0.3	
6	2.9	4.6	7.6	0.6	0.4	0.7	
7	0.8	1.7	2.6	0.4	0.6	0.6	
8	1.3	2.8	4.1	1.8	0.7	0.6	
9	3.4	8.0	11.9	0.8	0.4	2.5	
10	1.0	3.1	3.9	4.0	1.1	0.2	
11	0.6	1.3	2.0	0.3	0.4	0.3	

```
12                1.5                1.9       3.5      0.2      0.3      0.4
13                1.5                3.1       4.8      3.8      1.5      0.0
14                1.6                3.4       4.7      2.3      0.6      0.4
15                0.3                0.7       1.0      0.2      0.2      0.2
16                0.4                0.8       1.2      0.5      0.3      0.1
17                1.1                1.9       3.1      0.2      0.4      0.6
18                0.2                0.6       0.8      1.3      0.4      0.2
19                0.7                1.0       1.6      1.2      0.5      0.1

    Turnovers  TARGET_5Yrs
0         1.6            1
1         1.4            1
2         0.6            1
3         1.9            1
4         0.7            1
5         1.1            1
6         1.2            1
7         0.3            1
8         1.7            1
9         2.4            1
10        1.5            1
11        0.7            1
12        0.4            0
13        2.0            1
14        2.5            1
15        0.5            1
16        0.7            1
17        0.7            1
18        0.8            1
19        1.2            1

[20 rows x 21 columns]
```

[10]:
```
df_test = df_2022_test.rename(columns = {
    'GP':'Games Played',
    'MIN':'Minutes Played',
    'PTS':'Points Per Game',
    'FGM' : 'Field Goals Made',
    'FGA' : 'Field Goals Attempts',
    'FG%' : 'Field Goals Percent',
    '3P Made' : '3Points Made',
    '3PA' : '3Points Attempts',
    '3P%' : '3Points Percent',
    'FTM' : 'Free Throw Made',
    'FTA' : 'Free Throw Attempts',
    'FT%' : 'Free Throw Percent',
    'OREB' : 'Offensive Rebounds',
```

```
    'DREB' : 'Defensive Rebounds',
    'REB' : 'Rebounds',
    'AST' : 'Assists',
    'STL' : 'Steals',
    'BLK' : 'Blocks',
    'TOV' : 'Turnovers'
    })

df_test.head(20)
```

[10]:

	Id	Games Played	Minutes Played	Points Per Game	Field Goals Made	Field Goals Attempts	Field Goals Percent	3Points Made	3Points Attempts
0	0	56	9.1	4.0	1.6	3.7	43.7	0.1	0.3
1	1	43	19.3	10.1	3.7	8.1	46.0	0.6	1.7
2	2	82	33.9	11.3	4.9	10.6	45.6	0.5	1.9
3	3	86	44.7	18.8	6.8	15.9	42.9	0.5	1.8
4	4	58	12.3	4.7	1.6	4.0	40.0	0.5	1.7
5	5	59	16.1	7.1	2.8	4.5	62.0	-0.1	-0.6
6	6	45	8.3	3.2	1.4	2.8	46.4	-0.5	-1.6
7	7	53	12.1	4.7	1.8	4.1	44.1	0.1	0.2
8	8	41	21.6	7.9	3.1	6.0	49.6	0.3	0.9
9	9	25	10.0	2.6	1.0	2.3	45.0	0.2	0.7
10	10	72	12.6	3.5	1.3	3.2	41.9	0.1	0.1
11	11	67	21.3	11.1	4.6	8.3	53.9	0.0	0.2
12	12	47	31.1	11.5	4.1	9.9	42.1	1.0	3.0
13	13	87	20.7	10.9	4.5	9.8	48.4	0.5	1.1
14	14	82	43.4	17.3	7.0				
15	15	68	42.4	23.1	10.0				
16	16	78	23.9	8.1	3.1				
17	17	52	21.2	9.7	3.8				
18	18	27	8.5	1.6	0.7				
19	19	77	22.2	6.9	3.0				

14	15.4	46.2	1.0	2.5
15	20.8	48.7	0.6	1.7
16	5.8	52.1	-0.1	-0.1
17	8.0	49.0	0.7	2.1
18	1.4	50.4	-0.2	-0.6
19	6.6	45.1	0.1	0.2

	3Points Percent	Free Throw Made	Free Throw Attempts	Free Throw Percent	\
0	7.3	0.7	1.2	63.4	
1	35.1	1.8	2.5	75.3	
2	44.8	1.8	2.7	71.2	
3	13.5	4.5	6.3	70.9	
4	38.7	1.1	1.3	76.9	
5	3.4	1.7	2.7	58.6	
6	8.9	0.5	0.7	69.3	
7	25.7	0.9	1.0	80.2	
8	-14.8	1.8	2.2	77.6	
9	11.7	0.4	0.5	84.1	
10	14.9	0.5	0.6	74.8	
11	22.5	1.9	2.4	81.3	
12	28.7	2.1	2.6	72.3	
13	28.1	1.4	2.3	61.2	
14	31.2	2.4	2.9	84.3	
15	7.0	3.4	5.1	69.4	
16	33.8	2.1	3.2	64.3	
17	35.6	1.6	2.2	67.8	
18	22.4	0.1	0.2	55.8	
19	-8.2	1.0	1.6	49.2	

	Offensive Rebounds	Defensive Rebounds	Rebounds	Assists	Steals	Blocks	\
0	1.2	0.8	1.7	0.4	0.2	0.3	
1	0.5	0.9	1.5	3.5	0.6	0.0	
2	1.3	3.3	4.5	2.5	1.3	0.3	
3	1.5	3.2	5.0	4.1	0.9	0.1	
4	0.2	0.6	0.9	1.5	0.5	-0.4	
5	1.8	2.6	4.6	0.6	0.6	0.4	
6	0.5	0.5	1.2	0.2	0.1	0.2	
7	0.6	1.1	1.6	0.8	0.4	-0.2	
8	2.4	2.6	5.1	1.4	0.5	0.4	
9	0.4	1.5	1.8	0.4	0.2	0.2	
10	0.5	1.4	1.8	1.1	0.4	0.2	
11	1.6	2.4	4.0	3.3	1.3	0.3	
12	1.0	3.3	4.4	1.9	0.9	0.2	
13	0.9	1.4	2.2	3.2	1.0	0.0	
14	2.1	3.5	5.3	4.3	1.4	0.2	
15	2.0	4.3	6.4	4.2	1.7	0.3	
16	1.4	2.3	3.7	1.4	0.6	0.3	

```
17                1.0                2.5       3.4       1.7       0.9       0.3
18                0.6                1.1       1.5       0.4       0.3       0.0
19                2.2                2.3       4.1       1.4       0.4       0.6
```

```
    Turnovers
0         0.8
1         1.8
2         2.0
3         3.6
4         0.9
5         0.9
6         0.4
7         0.6
8         1.4
9         0.5
10        0.8
11        1.7
12        1.2
13        2.3
14        3.0
15        2.3
16        1.8
17        1.5
18        0.4
19        1.5
```

3 2. Quality investigation

3.1 2.1 Duplicates

[11]:
```
# Check number of duplicates
duplicates_count = df_main.duplicated().sum()
print(f"You have {duplicates_count} duplicates in the main data.")
```

```
You have 0 duplicates in the main data.
```

[12]:
```
# Check number of duplicates while ignoring the index feature
duplicates_count = df_main.drop(labels=["Id"], axis=1).duplicated().sum()
print(f"You have {duplicates_count} duplicates in the main data - ignoring the␣
  ↪Id.")
```

```
You have 0 duplicates in the main data - ignoring the Id.
```

[13]:
```
# Check number of duplicates
duplicates_count = df_test.duplicated().sum()
print(f"You have {duplicates_count} duplicates in the test data.")
```

```
You have 0 duplicates in the test data.
```

[14]:
```
# Check number of duplicates while ignoring the index feature
duplicates_count = df_test.drop(labels=["Id"], axis=1).duplicated().sum()
print(f"You have {duplicates_count} duplicates in the test data - ignoring the Id.")
```

```
You have 0 duplicates in the test data - ignoring the Id.
```

3.2 2.2 Missing values

[15]:
```
df_main.describe()
```

[15]:

	Id	Games Played	Minutes Played	Points Per Game	\
count	8000.00000	8000.000000	8000.000000	8000.000000	
mean	7798.50000	62.777875	18.576662	7.267088	
std	2309.54541	17.118774	8.935263	4.318732	
min	3799.00000	-8.000000	2.900000	0.800000	
25%	5798.75000	51.000000	12.000000	4.100000	
50%	7798.50000	63.000000	16.800000	6.300000	
75%	9798.25000	74.000000	23.500000	9.500000	
max	11798.00000	123.000000	73.800000	34.200000	

	Field Goals Made	Field Goals Attempts	Field Goals Percent	\
count	8000.000000	8000.000000	8000.000000	
mean	2.807037	6.231212	44.608900	
std	1.693373	3.584559	6.155453	
min	0.300000	0.800000	21.300000	
25%	1.600000	3.600000	40.400000	
50%	2.400000	5.400000	44.400000	
75%	3.700000	8.100000	48.700000	
max	13.100000	28.900000	67.200000	

	3Points Made	3Points Attempts	3Points Percent	…	\
count	8000.000000	8000.000000	8000.000000	…	
mean	0.264525	0.816562	19.583700	…	
std	0.384093	1.060964	16.003155	…	
min	-1.100000	-3.100000	-38.500000	…	
25%	0.000000	0.100000	8.400000	…	
50%	0.300000	0.800000	19.500000	…	
75%	0.500000	1.500000	30.600000	…	
max	1.700000	4.700000	82.100000	…	

	Free Throw Attempts	Free Throw Percent	Offensive Rebounds	\
count	8000.000000	8000.000000	8000.000000	
mean	1.947788	71.365825	1.077838	
std	1.252352	10.430447	0.785670	

```
min               0.000000          -13.300000            0.000000
25%               1.000000           65.000000            0.500000
50%               1.700000           71.400000            0.900000
75%               2.600000           77.500000            1.500000
max              11.100000          168.900000            5.500000

       Defensive Rebounds     Rebounds      Assists       Steals       Blocks  \
count         8000.000000  8000.000000  8000.000000  8000.000000  8000.000000
mean             2.168500     3.245300     1.624513     0.648687     0.245212
std              1.392224     2.085154     1.355986     0.407626     0.821037
min              0.200000     0.300000     0.000000     0.000000   -17.900000
25%              1.100000     1.700000     0.700000     0.300000     0.100000
50%              1.900000     2.800000     1.300000     0.600000     0.200000
75%              2.900000     4.300000     2.200000     0.900000     0.400000
max             11.000000    15.900000    12.800000     3.600000    18.900000

         Turnovers  TARGET_5Yrs
count  8000.000000  8000.000000
mean      1.257763     0.833625
std       0.723270     0.372440
min       0.100000     0.000000
25%       0.700000     1.000000
50%       1.100000     1.000000
75%       1.600000     1.000000
max       5.300000     1.000000

[8 rows x 21 columns]
```

[16]:
```
df_main.describe()
```

[16]:
```
                Id  Games Played  Minutes Played  Points Per Game  \
count   8000.00000   8000.000000     8000.000000      8000.000000
mean    7798.50000     62.777875       18.576662         7.267088
std     2309.54541     17.118774        8.935263         4.318732
min     3799.00000     -8.000000        2.900000         0.800000
25%     5798.75000     51.000000       12.000000         4.100000
50%     7798.50000     63.000000       16.800000         6.300000
75%     9798.25000     74.000000       23.500000         9.500000
max    11798.00000    123.000000       73.800000        34.200000

       Field Goals Made  Field Goals Attempts  Field Goals Percent  \
count       8000.000000           8000.000000          8000.000000
mean           2.807037              6.231212            44.608900
std            1.693373              3.584559             6.155453
min            0.300000              0.800000            21.300000
25%            1.600000              3.600000            40.400000
50%            2.400000              5.400000            44.400000
```

```
75%            3.700000              8.100000            48.700000
max           13.100000             28.900000            67.200000

       3Points Made  3Points Attempts  3Points Percent  …  \
count   8000.000000       8000.000000      8000.000000  …
mean       0.264525          0.816562        19.583700  …
std        0.384093          1.060964        16.003155  …
min       -1.100000         -3.100000       -38.500000  …
25%        0.000000          0.100000         8.400000  …
50%        0.300000          0.800000        19.500000  …
75%        0.500000          1.500000        30.600000  …
max        1.700000          4.700000        82.100000  …

       Free Throw Attempts  Free Throw Percent  Offensive Rebounds  \
count          8000.000000         8000.000000         8000.000000
mean              1.947788           71.365825            1.077838
std               1.252352           10.430447            0.785670
min               0.000000          -13.300000            0.000000
25%               1.000000           65.000000            0.500000
50%               1.700000           71.400000            0.900000
75%               2.600000           77.500000            1.500000
max              11.100000          168.900000            5.500000

       Defensive Rebounds     Rebounds      Assists       Steals       Blocks  \
count         8000.000000  8000.000000  8000.000000  8000.000000  8000.000000
mean             2.168500     3.245300     1.624513     0.648687     0.245212
std              1.392224     2.085154     1.355986     0.407626     0.821037
min              0.200000     0.300000     0.000000     0.000000   -17.900000
25%              1.100000     1.700000     0.700000     0.300000     0.100000
50%              1.900000     2.800000     1.300000     0.600000     0.200000
75%              2.900000     4.300000     2.200000     0.900000     0.400000
max             11.000000    15.900000    12.800000     3.600000    18.900000

         Turnovers  TARGET_5Yrs
count  8000.000000  8000.000000
mean      1.257763     0.833625
std       0.723270     0.372440
min       0.100000     0.000000
25%       0.700000     1.000000
50%       1.100000     1.000000
75%       1.600000     1.000000
max       5.300000     1.000000

[8 rows x 21 columns]
```

[17]:
```
df_test.describe()
```

[17]:

	Id	Games Played	Minutes Played	Points Per Game
count	3799.000000	3799.000000	3799.000000	3799.000000
mean	1899.000000	62.853909	18.650224	7.328034
std	1096.821164	17.151740	8.727259	4.294724
min	0.000000	6.000000	3.700000	0.700000
25%	949.500000	51.000000	12.200000	4.200000
50%	1899.000000	63.000000	17.000000	6.400000
75%	2848.500000	74.000000	23.300000	9.400000
max	3798.000000	126.000000	68.000000	33.000000

	Field Goals Made	Field Goals Attempts	Field Goals Percent
count	3799.000000	3799.000000	3799.000000
mean	2.835404	6.302580	44.599079
std	1.688427	3.579221	6.040168
min	0.300000	0.800000	25.100000
25%	1.600000	3.700000	40.500000
50%	2.500000	5.500000	44.600000
75%	3.700000	8.100000	48.500000
max	13.400000	26.200000	74.600000

	3Points Made	3Points Attempts	3Points Percent	Free Throw Made
count	3799.000000	3799.000000	3799.000000	3799.000000
mean	0.255962	0.796920	19.234746	1.399842
std	0.380987	1.052862	15.968989	0.926140
min	-1.000000	-2.700000	-38.000000	0.000000
25%	0.000000	0.100000	8.500000	0.700000
50%	0.300000	0.800000	19.400000	1.200000
75%	0.500000	1.500000	30.250000	1.900000
max	1.600000	4.300000	73.800000	7.800000

	Free Throw Attempts	Free Throw Percent	Offensive Rebounds
count	3799.000000	3799.000000	3799.000000
mean	1.953567	71.612924	1.096025
std	1.250376	10.457336	0.785678
min	0.000000	23.700000	0.000000
25%	1.000000	65.000000	0.500000
50%	1.700000	71.500000	0.900000
75%	2.600000	78.000000	1.500000
max	9.800000	127.100000	6.900000

	Defensive Rebounds	Rebounds	Assists	Steals	Blocks
count	3799.000000	3799.000000	3799.000000	3799.000000	3799.000000
mean	2.179495	3.275783	1.636483	0.653593	0.257726
std	1.371935	2.070646	1.335496	0.410573	0.639660
min	0.200000	0.300000	0.000000	0.000000	-7.100000
25%	1.200000	1.800000	0.600000	0.400000	0.100000
50%	1.900000	2.800000	1.300000	0.600000	0.200000

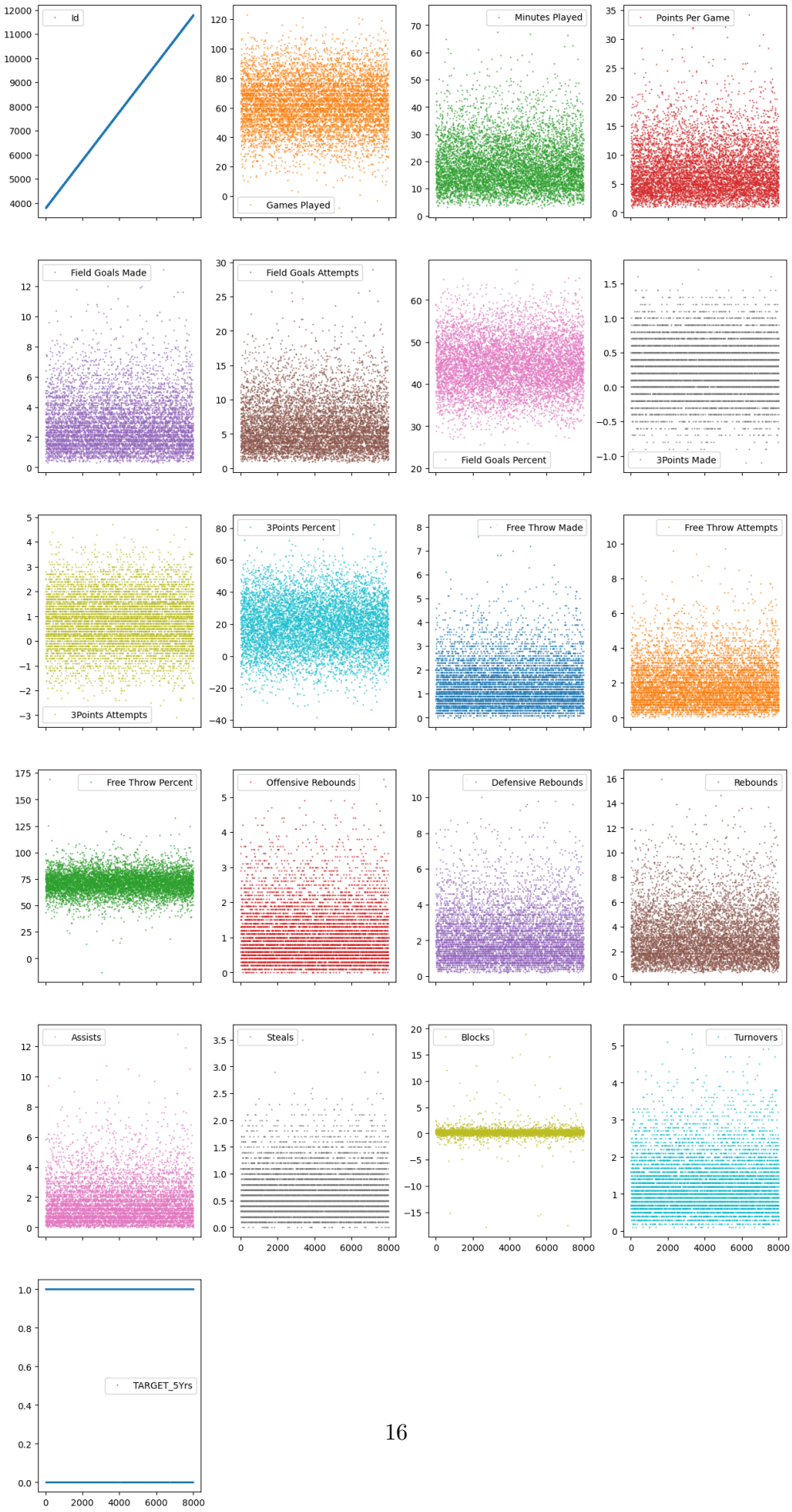
```
75%              2.900000     4.300000     2.300000     0.900000     0.400000
max             12.000000    18.500000     9.000000     2.700000    14.800000

         Turnovers
count  3799.000000
mean      1.257910
std       0.712449
min       0.100000
25%       0.700000
50%       1.100000
75%       1.600000
max       5.200000
```

[18]:
```
df_test.describe()
```

[18]:
```
                Id  Games Played  Minutes Played  Points Per Game  \
count  3799.000000   3799.000000     3799.000000      3799.000000
mean   1899.000000     62.853909       18.650224         7.328034
std    1096.821164     17.151740        8.727259         4.294724
min       0.000000      6.000000        3.700000         0.700000
25%     949.500000     51.000000       12.200000         4.200000
50%    1899.000000     63.000000       17.000000         6.400000
75%    2848.500000     74.000000       23.300000         9.400000
max    3798.000000    126.000000       68.000000        33.000000

       Field Goals Made  Field Goals Attempts  Field Goals Percent  \
count       3799.000000           3799.000000          3799.000000
mean           2.835404              6.302580            44.599079
std            1.688427              3.579221             6.040168
min            0.300000              0.800000            25.100000
25%            1.600000              3.700000            40.500000
50%            2.500000              5.500000            44.600000
75%            3.700000              8.100000            48.500000
max           13.400000             26.200000            74.600000

       3Points Made  3Points Attempts  3Points Percent  Free Throw Made  \
count   3799.000000       3799.000000      3799.000000      3799.000000
mean       0.255962          0.796920        19.234746         1.399842
std        0.380987          1.052862        15.968989         0.926140
min       -1.000000         -2.700000       -38.000000         0.000000
25%        0.000000          0.100000         8.500000         0.700000
50%        0.300000          0.800000        19.400000         1.200000
75%        0.500000          1.500000        30.250000         1.900000
max        1.600000          4.300000        73.800000         7.800000

       Free Throw Attempts  Free Throw Percent  Offensive Rebounds  \
count          3799.000000         3799.000000         3799.000000
```

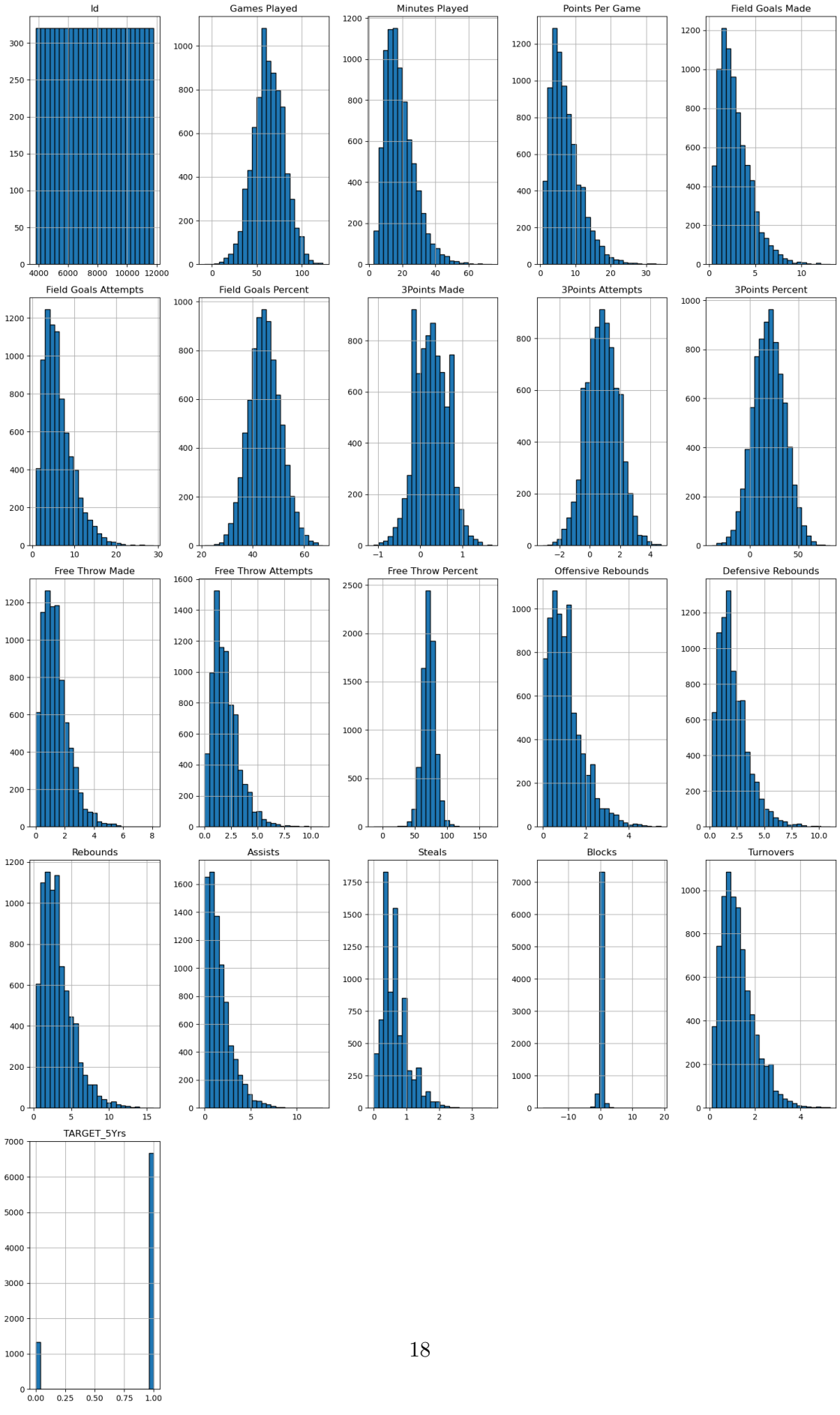
```
mean              1.953567          71.612924           1.096025
std               1.250376          10.457336           0.785678
min               0.000000          23.700000           0.000000
25%               1.000000          65.000000           0.500000
50%               1.700000          71.500000           0.900000
75%               2.600000          78.000000           1.500000
max               9.800000         127.100000           6.900000

       Defensive Rebounds     Rebounds      Assists       Steals       Blocks  \
count         3799.000000  3799.000000  3799.000000  3799.000000  3799.000000
mean             2.179495     3.275783     1.636483     0.653593     0.257726
std              1.371935     2.070646     1.335496     0.410573     0.639660
min              0.200000     0.300000     0.000000     0.000000    -7.100000
25%              1.200000     1.800000     0.600000     0.400000     0.100000
50%              1.900000     2.800000     1.300000     0.600000     0.200000
75%              2.900000     4.300000     2.300000     0.900000     0.400000
max             12.000000    18.500000     9.000000     2.700000    14.800000

         Turnovers
count  3799.000000
mean      1.257910
std       0.712449
min       0.100000
25%       0.700000
50%       1.100000
75%       1.600000
max       5.200000
```

[19]:
```
df_main.plot(lw=0, marker=".", subplots=True, layout=(-1, 4),
          figsize=(15, 30), markersize=1);
```


Id
Games Played
Minutes Played
Points Per Game
Field Goals Made
Field Goals Attempts
Field Goals Percent
3Points Made
3Points Attempts
3Points Percent
Free Throw Made
Free Throw Attempts
Free Throw Percent
Offensive Rebounds
Defensive Rebounds
Rebounds
Assists
Steals
Blocks
Turnovers
TARGET_5Yrs

4 3. Content investigation

[20]:
```
# using matplotlib.pyplot - plt

# Plots the histogram for each numerical feature in a separate subplot
df_main.hist(bins=25, figsize=(15, 25), layout=(-1, 5), edgecolor="black")
plt.tight_layout();
```

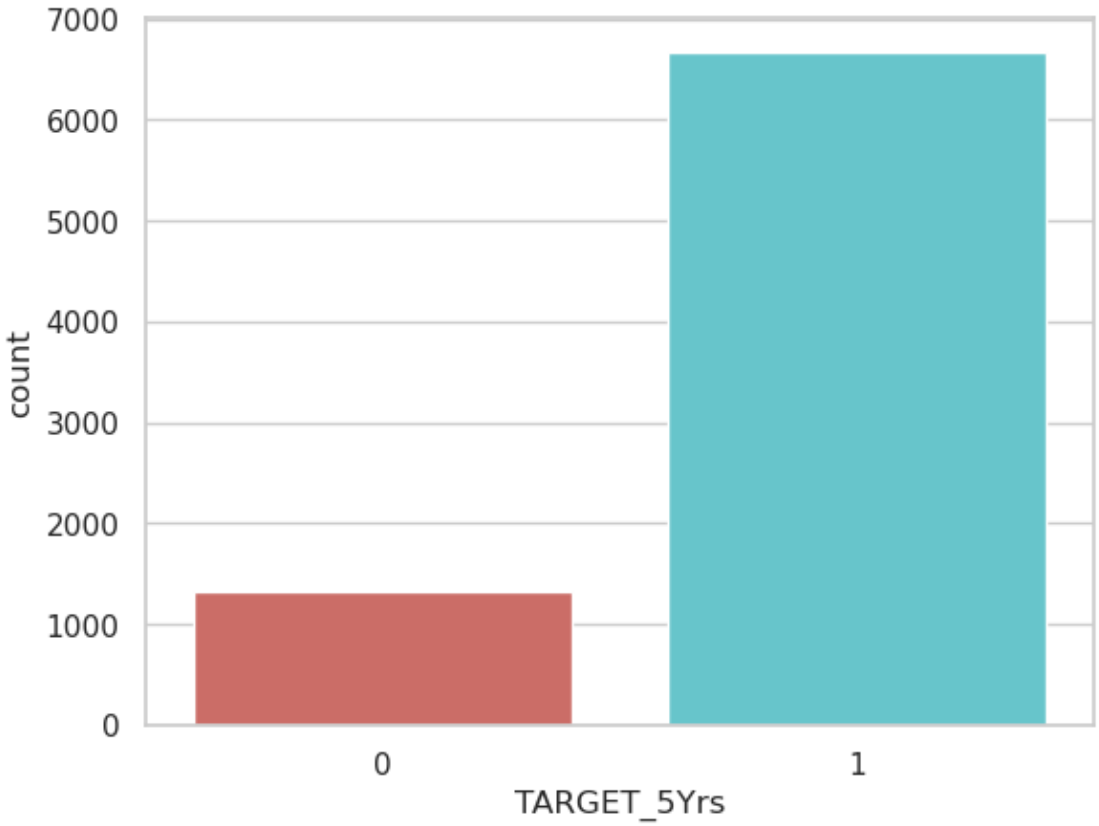

Id
Games Played
Minutes Played
Points Per Game
Field Goals Made
Field Goals Attempts
Field Goals Percent
3Points Made
3Points Attempts
3Points Percent
Free Throw Made
Free Throw Attempts
Free Throw Percent
Offensive Rebounds
Defensive Rebounds
Rebounds
Assists
Steals
Blocks
Turnovers
TARGET_5Yrs

5 4. Data Exploration

[21]:
```
df_main['TARGET_5Yrs'].value_counts()
```

[21]:
```
1    6669
0    1331
Name: TARGET_5Yrs, dtype: int64
```

[22]:
```
# using seaborn - sns

sns.set(style="white")
sns.set(style="whitegrid", color_codes=True)

sns.countplot(x = 'TARGET_5Yrs',
              data = df_main,
              palette = 'hls'
             )
plt.show()
plt.savefig('count_plot')
```



```
<Figure size 640x480 with 0 Axes>
```

[26]:
```
count_lt_5Yrs = len(df_main[df_main['TARGET_5Yrs']==0])
count_gt_5Yrs = len(df_main[df_main['TARGET_5Yrs']==1])
pct_of_lt_5Yrs = count_lt_5Yrs/(count_lt_5Yrs+count_gt_5Yrs)
print("Percentage of NBA career less than 5 years is", pct_of_lt_5Yrs*100)
pct_of_gt_5Yrs = count_gt_5Yrs/(count_lt_5Yrs+count_gt_5Yrs)
print("Percentage of NBA career at least 5 years is", pct_of_gt_5Yrs*100)
```

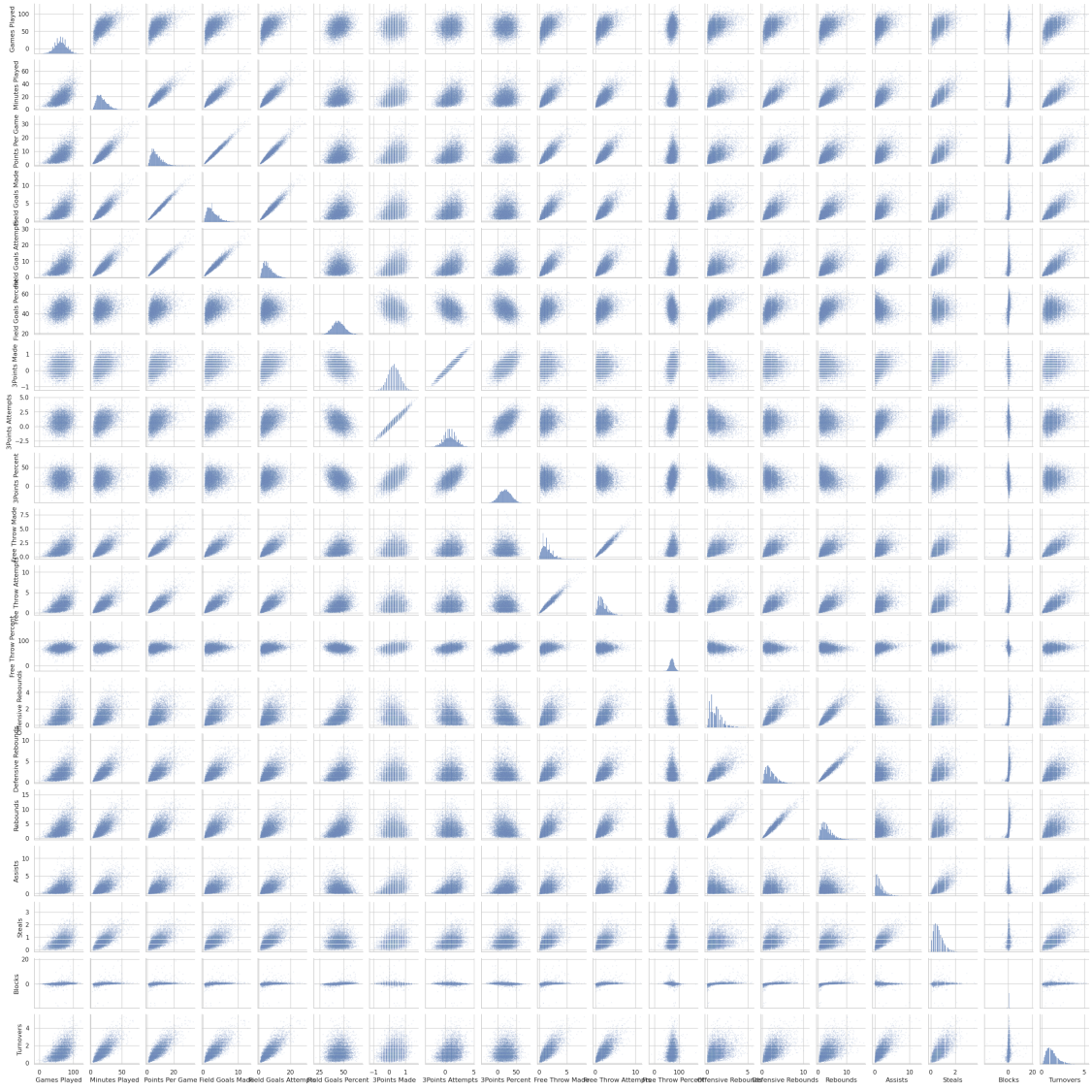
```
Percentage of NBA career less than 5 years is 16.6375
Percentage of NBA career at least 5 years is 83.3625
```

[27]:
```
df_main.groupby('TARGET_5Yrs').mean()
```

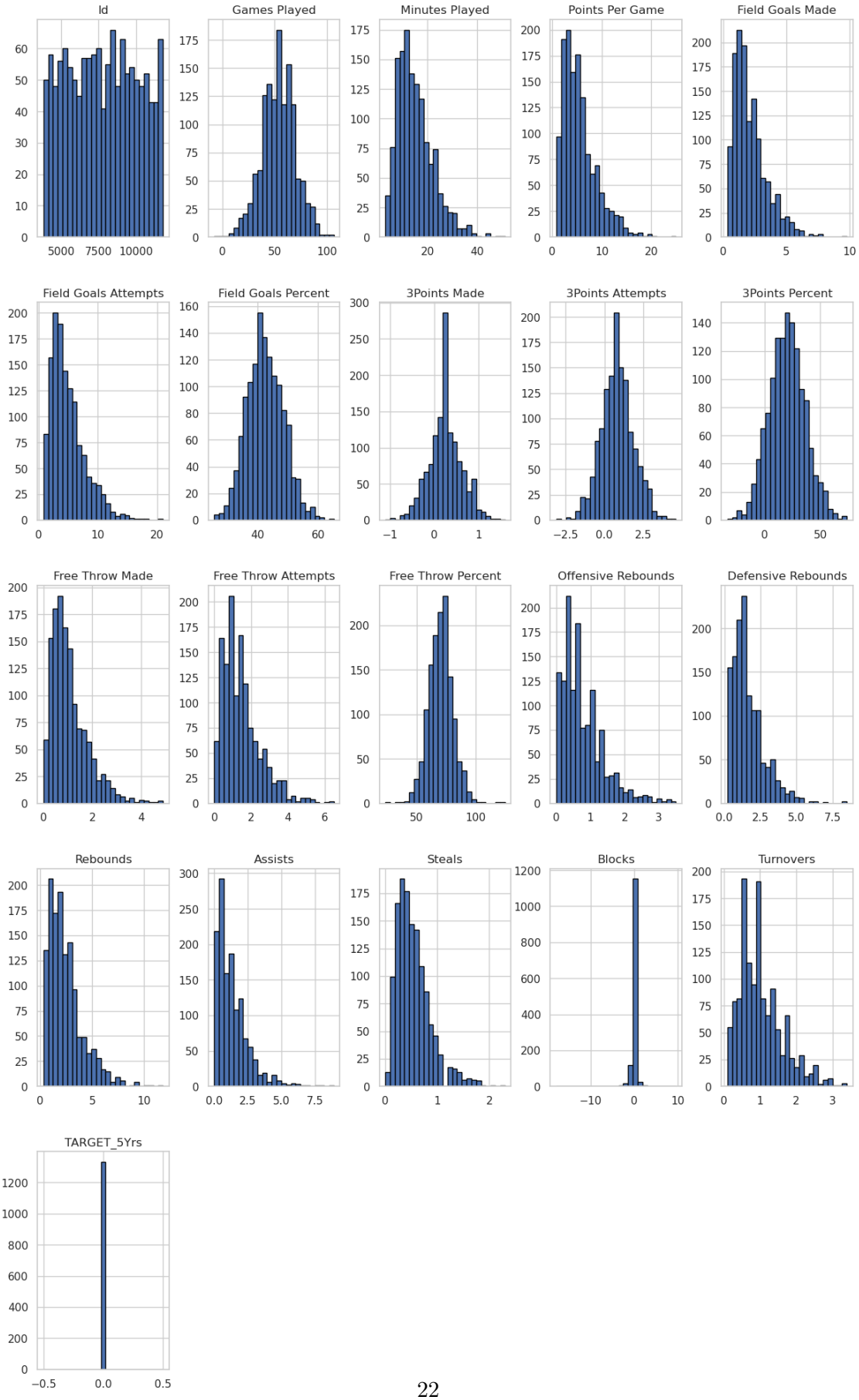
[27]:

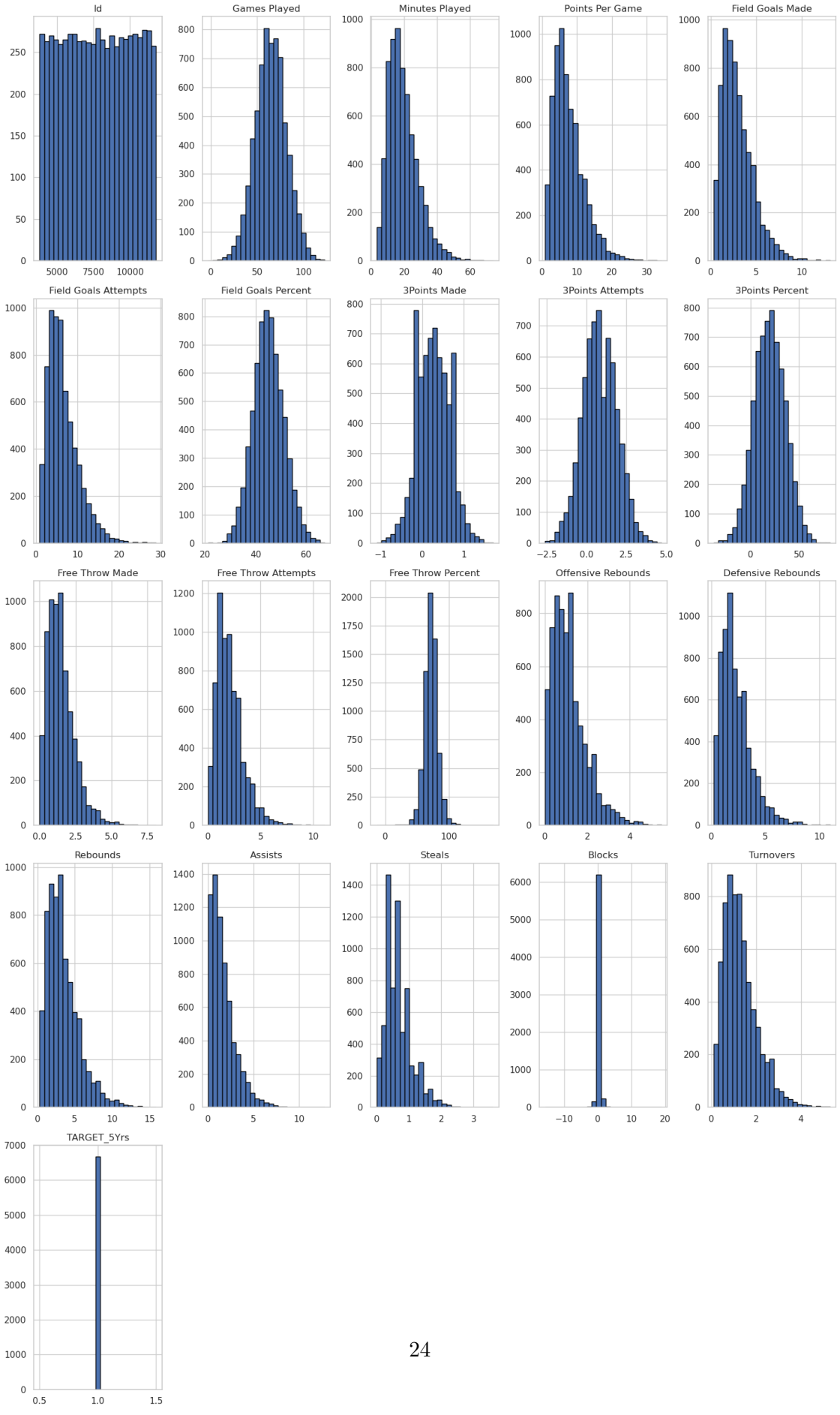
TARGET_5Yrs	Id	Games Played	Minutes Played	Points Per Game	Field Goals Made	Field Goals Attempts	Field Goals Percent	3Points Made	3Points Attempts	3Points Percent	Free Throw Made	Free Throw Attempts	Free Throw Percent	Offensive Rebounds	Defensive Rebounds	Rebounds	Assists	Steals	Blocks	Turnovers
0	7764.149512	53.501127	14.932682	5.507739	2.111270	4.928325	42.420210	0.251615	0.808340	20.012697	1.032006	1.452968	70.445304	0.774305	1.644778	2.425620	1.330804	0.531555	0.063937	1.012923
1	7805.355676	64.629330	19.303929	7.618219	2.945899	6.491243	45.045719	0.267102	0.818204	19.498081	1.464477	2.046544	71.549543	1.138417	2.273024	3.408892	1.683131	0.672065	0.281392	1.306628

[31]:
```
df_temp = df_main.drop(columns=['Id', 'TARGET_5Yrs'], axis=1)

sns.pairplot(df_temp,
             height=1.5,
             plot_kws={"s":2, "alpha":0.2}
            );
```



[29]:
```
# using matplotlib.pyplot - plt

# Plots the histogram for each numerical feature in a separate subplot
df_main.groupby('TARGET_5Yrs').hist(bins=25, figsize=(15, 25), layout=(-1, 5),␣
  ↪edgecolor="black")
plt.tight_layout();
```


Id
Games Played
Minutes Played
Points Per Game
Field Goals Made
Field Goals Attempts
Field Goals Percent
3Points Made
3Points Attempts
3Points Percent
Free Throw Made
Free Throw Attempts
Free Throw Percent
Offensive Rebounds
Defensive Rebounds
Rebounds
Assists
Steals
Blocks
Turnovers
TARGET_5Yrs


Id
Games Played
Minutes Played
Points Per Game
Field Goals Made
Field Goals Attempts
Field Goals Percent
3Points Made
3Points Attempts
3Points Percent
Free Throw Made
Free Throw Attempts
Free Throw Percent
Offensive Rebounds
Defensive Rebounds
Rebounds
Assists
Steals
Blocks
Turnovers
TARGET_5Yrs

[ ]: