In [57]:
```
import pandas as pd
```

In [59]:
```
teams = pd.read_csv("C:\\Users\\Admin\\Downloads\\teams.csv")
```

In [60]:
```
teams
```

Out[60]:

	team	country	year	events	athletes	age	height	weight	medals	prev_medals	prev_3_r
0	AFG	Afghanistan	1964	8	8	22.0	161.0	64.2	0	0.0	
1	AFG	Afghanistan	1968	5	5	23.2	170.2	70.0	0	0.0	
2	AFG	Afghanistan	1972	8	8	29.0	168.3	63.8	0	0.0	
3	AFG	Afghanistan	1980	11	11	23.6	168.4	63.2	0	0.0	
4	AFG	Afghanistan	2004	5	5	18.6	170.8	64.8	0	0.0	
...	...	...	...	...	...	...	...	...	...	...	
2139	ZIM	Zimbabwe	2000	19	26	25.0	179.0	71.1	0	0.0	
2140	ZIM	Zimbabwe	2004	11	14	25.1	177.8	70.5	3	0.0	
2141	ZIM	Zimbabwe	2008	15	16	26.1	171.9	63.7	4	3.0	
2142	ZIM	Zimbabwe	2012	8	9	27.3	174.4	65.2	0	4.0	
2143	ZIM	Zimbabwe	2016	13	31	27.5	167.8	62.2	0	0.0	

2144 rows × 11 columns

In [66]:
```
teams = teams[["team", "country", "year", "athletes", "age", "prev_medals", "medals"]]
```

In [70]:
```
teams
```

Out[70]:

	team	country	year	athletes	age	prev_medals	medals
0	AFG	Afghanistan	1964	8	22.0	0.0	0
1	AFG	Afghanistan	1968	5	23.2	0.0	0
2	AFG	Afghanistan	1972	8	29.0	0.0	0
3	AFG	Afghanistan	1980	11	23.6	0.0	0
4	AFG	Afghanistan	2004	5	18.6	0.0	0
...	...	...	...	...	...	...	...
2139	ZIM	Zimbabwe	2000	26	25.0	0.0	0
2140	ZIM	Zimbabwe	2004	14	25.1	0.0	3
2141	ZIM	Zimbabwe	2008	16	26.1	3.0	4
2142	ZIM	Zimbabwe	2012	9	27.3	4.0	0
2143	ZIM	Zimbabwe	2016	31	27.5	0.0	0

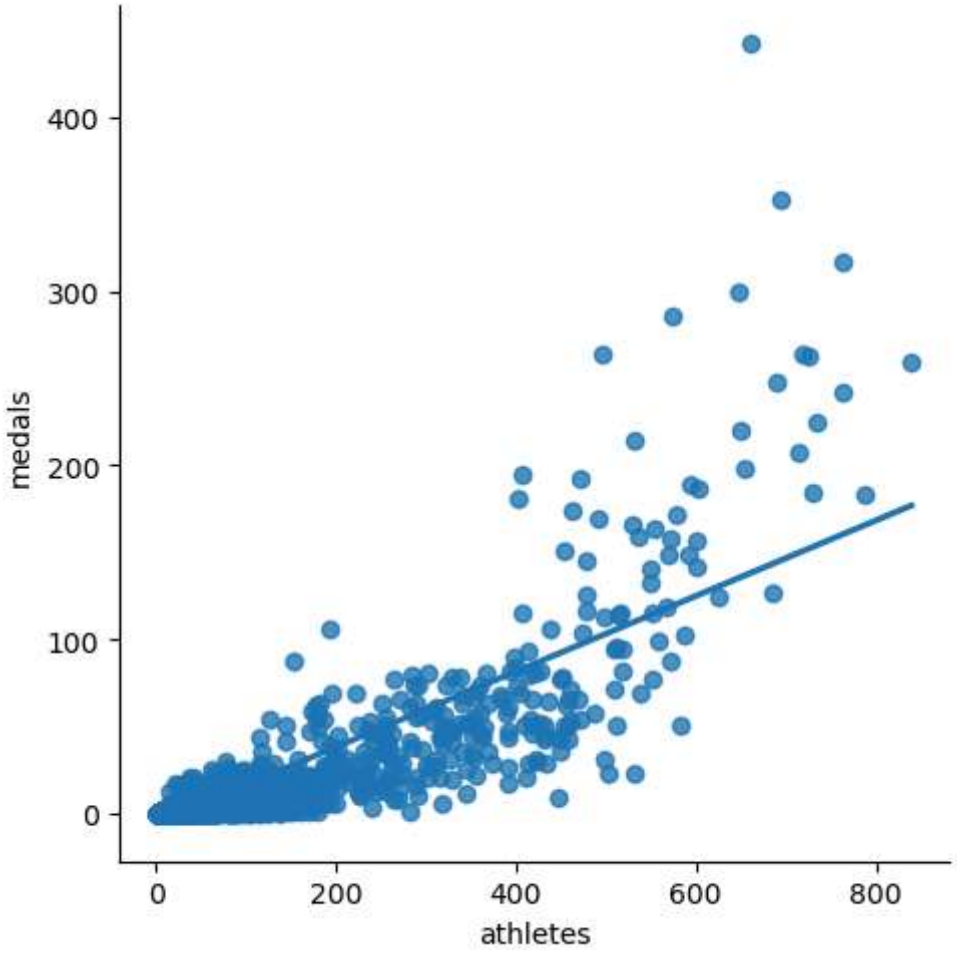
2144 rows × 7 columns

In [72]:
```
import seaborn as sns
```

In [73]:
```
sns.lmplot(x="athletes", y="medals", data=teams, fit_reg=True, ci=None)
```

```
C:\Users\Admin\anaconda3\Lib\site-packages\seaborn\axisgrid.py:118: UserWarning: The
figure layout has changed to tight
  self._figure.tight_layout(*args, **kwargs)
```

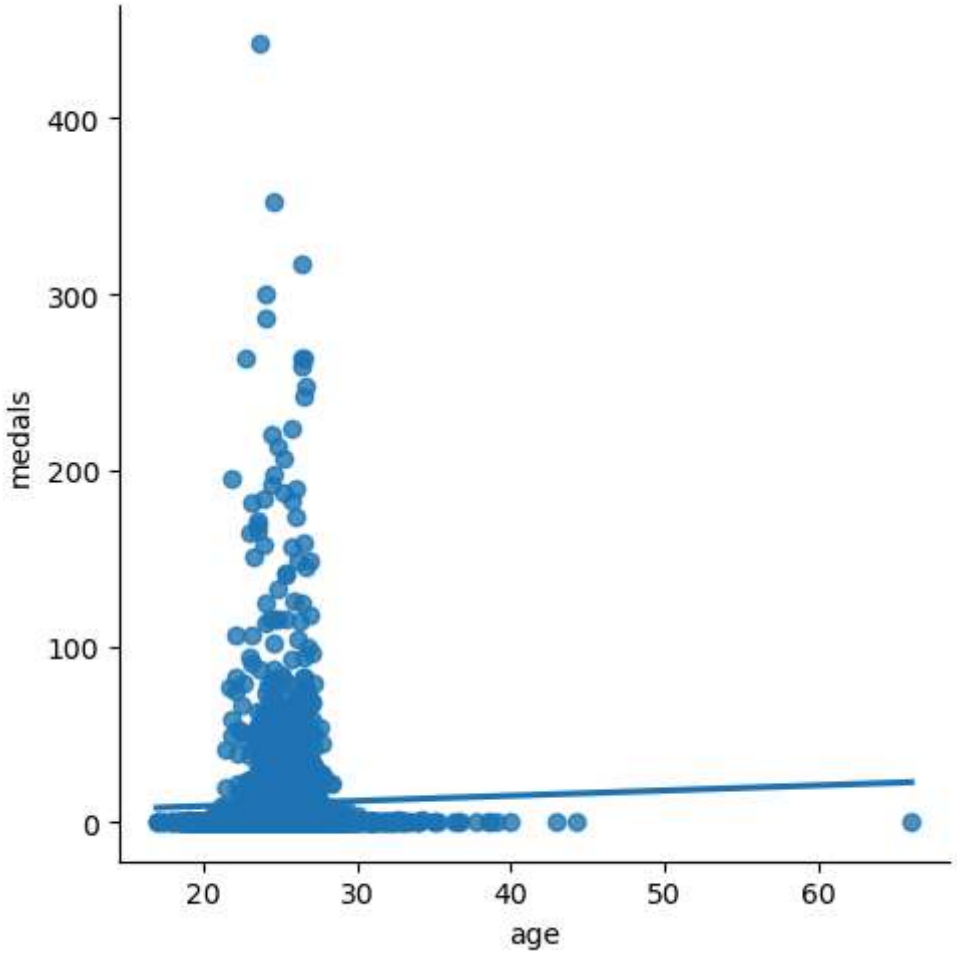
Out[73]:
```
<seaborn.axisgrid.FacetGrid at 0x177d34bc610>
```



In [75]:
```
sns.lmplot(x="age", y="medals", data=teams, fit_reg=True, ci=None)
```

```
C:\Users\Admin\anaconda3\Lib\site-packages\seaborn\axisgrid.py:118: UserWarning: The
figure layout has changed to tight
  self._figure.tight_layout(*args, **kwargs)
```

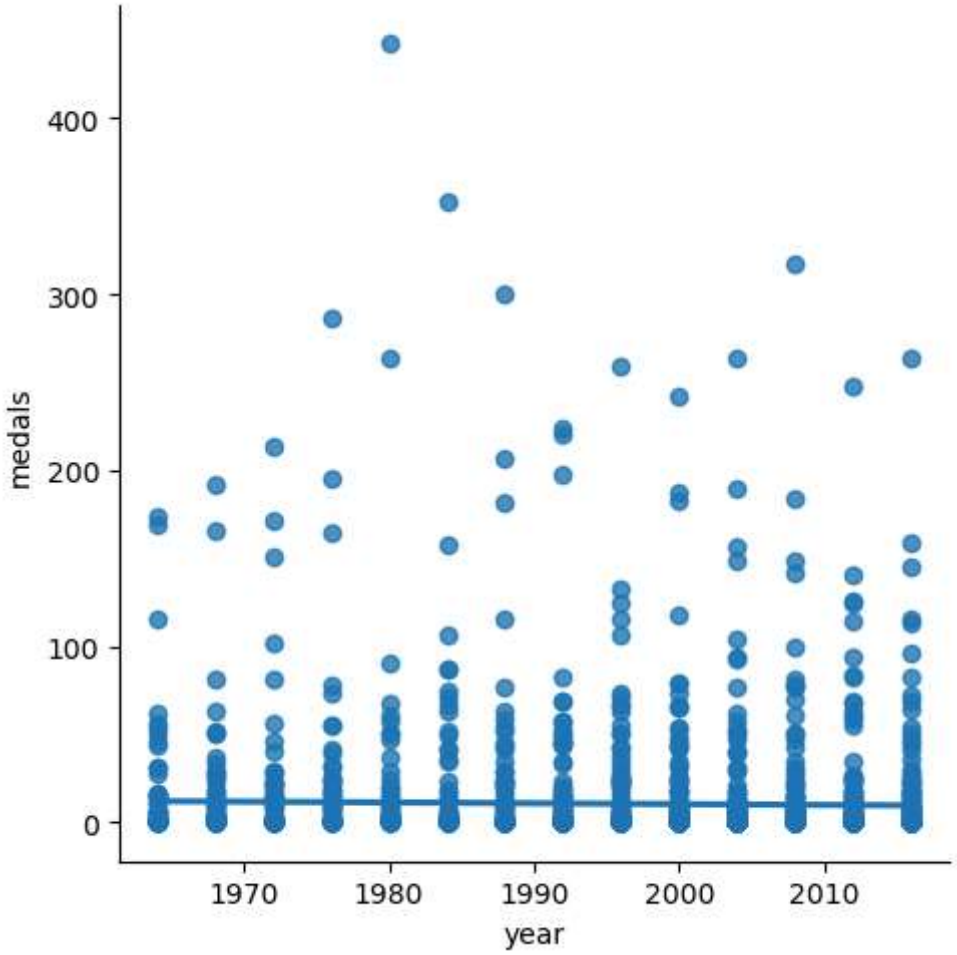
Out[75]:
```
<seaborn.axisgrid.FacetGrid at 0x177d3503f50>
```



In [76]:
```
sns.lmplot(x="year", y="medals", data=teams, fit_reg=True, ci=None)
```

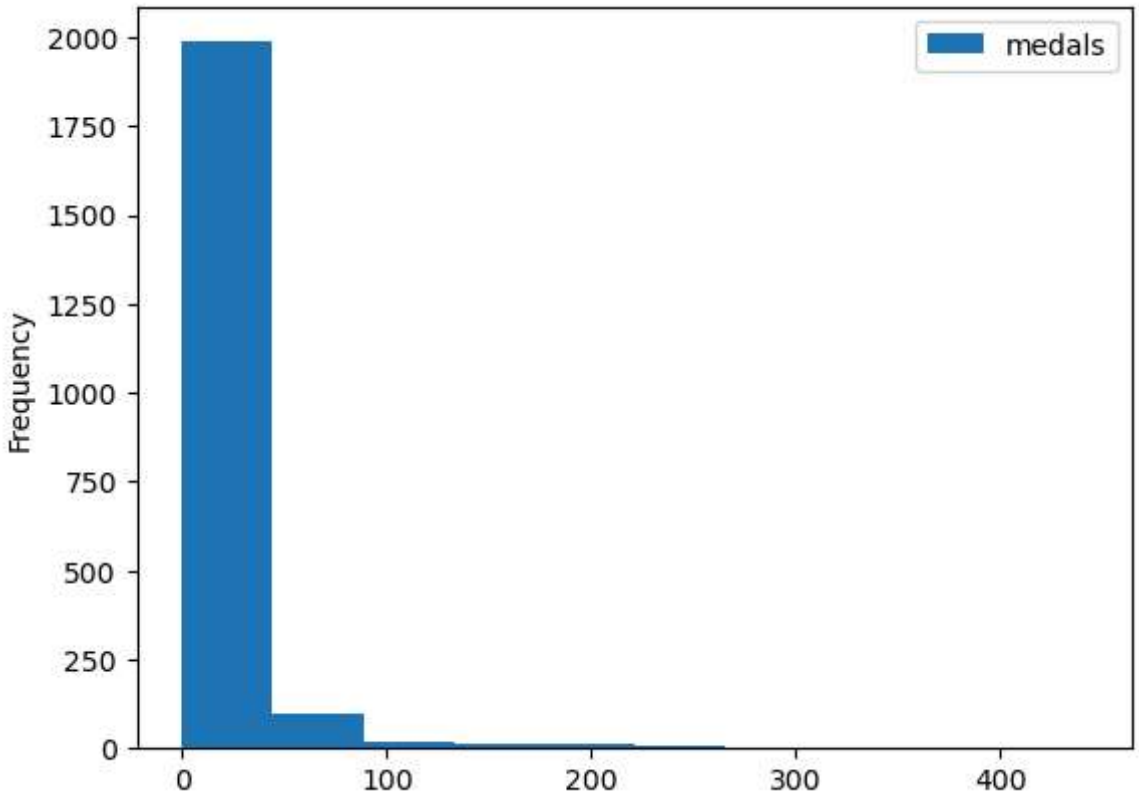
```
C:\Users\Admin\anaconda3\Lib\site-packages\seaborn\axisgrid.py:118: UserWarning: The
figure layout has changed to tight
  self._figure.tight_layout(*args, **kwargs)
```

Out[76]:
```
<seaborn.axisgrid.FacetGrid at 0x177d35555d0>
```



In [77]:
```
teams.plot.hist(y="medals")
```

Out[77]:
```
<Axes: ylabel='Frequency'>
```



In [78]:
```
teams[teams.isnull().any(axis=1)]
```

Out[78]:

	team	country	year	athletes	age	prev_medals	medals
19	ALB	Albania	1992	9	25.3	NaN	0
26	ALG	Algeria	1964	7	26.0	NaN	0
39	AND	Andorra	1976	3	28.3	NaN	0
50	ANG	Angola	1980	17	17.4	NaN	0
59	ANT	Antigua and Barbuda	1976	17	23.2	NaN	0
...	...	...	...	...	...	...	...
2092	VIN	Saint Vincent and the Grenadines	1988	6	20.5	NaN	0
2103	YAR	North Yemen	1984	3	27.7	NaN	0
2105	YEM	Yemen	1992	8	19.6	NaN	0
2112	YMD	South Yemen	1988	5	23.6	NaN	0
2120	ZAM	Zambia	1964	15	21.7	NaN	0

130 rows × 7 columns

In [81]:
```
teams = teams.dropna()
```

In [82]:
```
teams
```

Out[82]:

	team	country	year	athletes	age	prev_medals	medals
0	AFG	Afghanistan	1964	8	22.0	0.0	0
1	AFG	Afghanistan	1968	5	23.2	0.0	0
2	AFG	Afghanistan	1972	8	29.0	0.0	0
3	AFG	Afghanistan	1980	11	23.6	0.0	0
4	AFG	Afghanistan	2004	5	18.6	0.0	0
...	...	...	...	...	...	...	...
2139	ZIM	Zimbabwe	2000	26	25.0	0.0	0
2140	ZIM	Zimbabwe	2004	14	25.1	0.0	3
2141	ZIM	Zimbabwe	2008	16	26.1	3.0	4
2142	ZIM	Zimbabwe	2012	9	27.3	4.0	0
2143	ZIM	Zimbabwe	2016	31	27.5	0.0	0

2014 rows × 7 columns

In [87]:
```
train = teams[teams["year"] < 2012].copy()
```

In [88]:
```
train
```

Out[88]:

	team	country	year	athletes	age	prev_medals	medals
0	AFG	Afghanistan	1964	8	22.0	0.0	0
1	AFG	Afghanistan	1968	5	23.2	0.0	0
2	AFG	Afghanistan	1972	8	29.0	0.0	0
3	AFG	Afghanistan	1980	11	23.6	0.0	0
4	AFG	Afghanistan	2004	5	18.6	0.0	0
...	...	...	...	...	...	...	...
2137	ZIM	Zimbabwe	1992	28	21.2	0.0	0
2138	ZIM	Zimbabwe	1996	21	23.8	0.0	0
2139	ZIM	Zimbabwe	2000	26	25.0	0.0	0
2140	ZIM	Zimbabwe	2004	14	25.1	0.0	3
2141	ZIM	Zimbabwe	2008	16	26.1	3.0	4

1609 rows × 7 columns

In [89]:
```
test = teams[teams["year"] >= 2012].copy()
```

In [90]:
```
test
```

Out[90]:

	team	country	year	athletes	age	prev_medals	medals
6	AFG	Afghanistan	2012	6	24.8	1.0	1
7	AFG	Afghanistan	2016	3	24.7	1.0	0
24	ALB	Albania	2012	10	25.7	0.0	0
25	ALB	Albania	2016	6	23.7	0.0	0
37	ALG	Algeria	2012	39	24.8	2.0	1
...	...	...	...	...	...	...	...
2111	YEM	Yemen	2016	3	19.3	0.0	0
2131	ZAM	Zambia	2012	7	22.6	0.0	0
2132	ZAM	Zambia	2016	7	24.1	0.0	0
2142	ZIM	Zimbabwe	2012	9	27.3	4.0	0
2143	ZIM	Zimbabwe	2016	31	27.5	0.0	0

405 rows × 7 columns

In [91]:
```
train.shape
```

Out[91]:
```
(1609, 7)
```

In [92]:
```
test.shape
```

Out[92]:
```
(405, 7)
```

In [93]:
```
from sklearn.linear_model import LinearRegression
```

In [94]:
```
reg = LinearRegression()
```

In [95]:
```
predictors = ["athletes", "prev_medals"]
```

In [96]:
```
target = "medals"
```

In [97]:
```
reg.fit(train[predictors], train["medals"])
```

Out[97]:
```
▾ LinearRegression
LinearRegression()
```

In [98]:
```
predictions = reg.predict(test[predictors])
```

In [100…
```
test["predictions"] = predictions
```

In [101…
```
test
```

Out[101]:

	team	country	year	athletes	age	prev_medals	medals	predictions
6	AFG	Afghanistan	2012	6	24.8	1.0	1	-0.961221
7	AFG	Afghanistan	2016	3	24.7	1.0	0	-1.176333
24	ALB	Albania	2012	10	25.7	0.0	0	-1.425032
25	ALB	Albania	2016	6	23.7	0.0	0	-1.711847
37	ALG	Algeria	2012	39	24.8	2.0	1	2.155629
...	...	...	...	...	...	...	...	...
2111	YEM	Yemen	2016	3	19.3	0.0	0	-1.926958
2131	ZAM	Zambia	2012	7	22.6	0.0	0	-1.640143
2132	ZAM	Zambia	2016	7	24.1	0.0	0	-1.640143
2142	ZIM	Zimbabwe	2012	9	27.3	4.0	0	1.505767
2143	ZIM	Zimbabwe	2016	31	27.5	0.0	0	0.080748

405 rows × 8 columns

In [102…
```
test.loc[test["predictions"] < 0, "predictions"] = 0
```

In [103…
```
test["predictions"] = test["predictions"].round()
```

In [104…
```
test
```

Out[104]:

	team	country	year	athletes	age	prev_medals	medals	predictions
6	AFG	Afghanistan	2012	6	24.8	1.0	1	0.0
7	AFG	Afghanistan	2016	3	24.7	1.0	0	0.0
24	ALB	Albania	2012	10	25.7	0.0	0	0.0
25	ALB	Albania	2016	6	23.7	0.0	0	0.0
37	ALG	Algeria	2012	39	24.8	2.0	1	2.0
...	...	...	...	...	...	...	...	...
2111	YEM	Yemen	2016	3	19.3	0.0	0	0.0
2131	ZAM	Zambia	2012	7	22.6	0.0	0	0.0
2132	ZAM	Zambia	2016	7	24.1	0.0	0	0.0
2142	ZIM	Zimbabwe	2012	9	27.3	4.0	0	2.0
2143	ZIM	Zimbabwe	2016	31	27.5	0.0	0	0.0

405 rows × 8 columns

In [105…
```
from sklearn.metrics import mean_absolute_error
```

In [106…
```
error = mean_absolute_error(test["medals"], test["predictions"])
```

In [107…
```
error
```

Out[107]:
```
3.2987654320987656
```

In [108…
```
teams.describe()["medals"]
```

Out[108]:
```
count    2014.000000
mean       10.990070
std        33.627528
min         0.000000
25%         0.000000
50%         0.000000
75%         5.000000
max       442.000000
Name: medals, dtype: float64
```

In [109…
```
test[test["team"] == "USA"]
```

Out[109]:

	team	country	year	athletes	age	prev_medals	medals	predictions
2053	USA	United States	2012	689	26.7	317.0	248	285.0
2054	USA	United States	2016	719	26.4	248.0	264	236.0

In [110…
```
test[test["team"] == "IND"]
```

Out[110]:

	team	country	year	athletes	age	prev_medals	medals	predictions
907	IND	India	2012	95	26.0	3.0	6	7.0
908	IND	India	2016	130	26.1	6.0	2	12.0

In [116…
```
errors = (test["medals"] - test["predictions"]).abs()
```

In [117…
```
errors
```

Out[117]:
```
6       1.0
7       0.0
24      0.0
25      0.0
37      1.0
       ...
2111    0.0
2131    0.0
2132    0.0
2142    2.0
2143    0.0
Length: 405, dtype: float64
```

In [118…
```
errors_by_team = errors.groupby(test["team"]).mean()
```

In [119…
```
errors_by_team
```

Out[119]:
```
team
AFG    0.5
ALB    0.0
ALG    1.5
AND    0.0
ANG    0.0
      ...
VIE    1.0
VIN    0.0
YEM    0.0
ZAM    0.0
ZIM    1.0
Length: 204, dtype: float64
```

In [124…
```
medals_by_team = test["medals"].groupby(test["team"]).mean()
```

In [128…
```
error_ratio = errors_by_team / medals_by_team
```

In [129…
```
error_ratio
```

Out[129]:
```
team
AFG    1.0
ALB    NaN
ALG    1.0
AND    NaN
ANG    NaN
      ...
VIE    1.0
VIN    NaN
YEM    NaN
ZAM    NaN
ZIM    inf
Length: 204, dtype: float64
```

In [133…
```
error_ratio[~pd.isnull(error_ratio)]
```

Out[133]:
```
team
AFG    1.000000
ALG    1.000000
ARG    0.853659
ARM    0.428571
AUS    0.367347
         ...
USA    0.126953
UZB    0.625000
VEN    1.750000
VIE    1.000000
ZIM         inf
Length: 102, dtype: float64
```

In [138…
```
import numpy as np
```

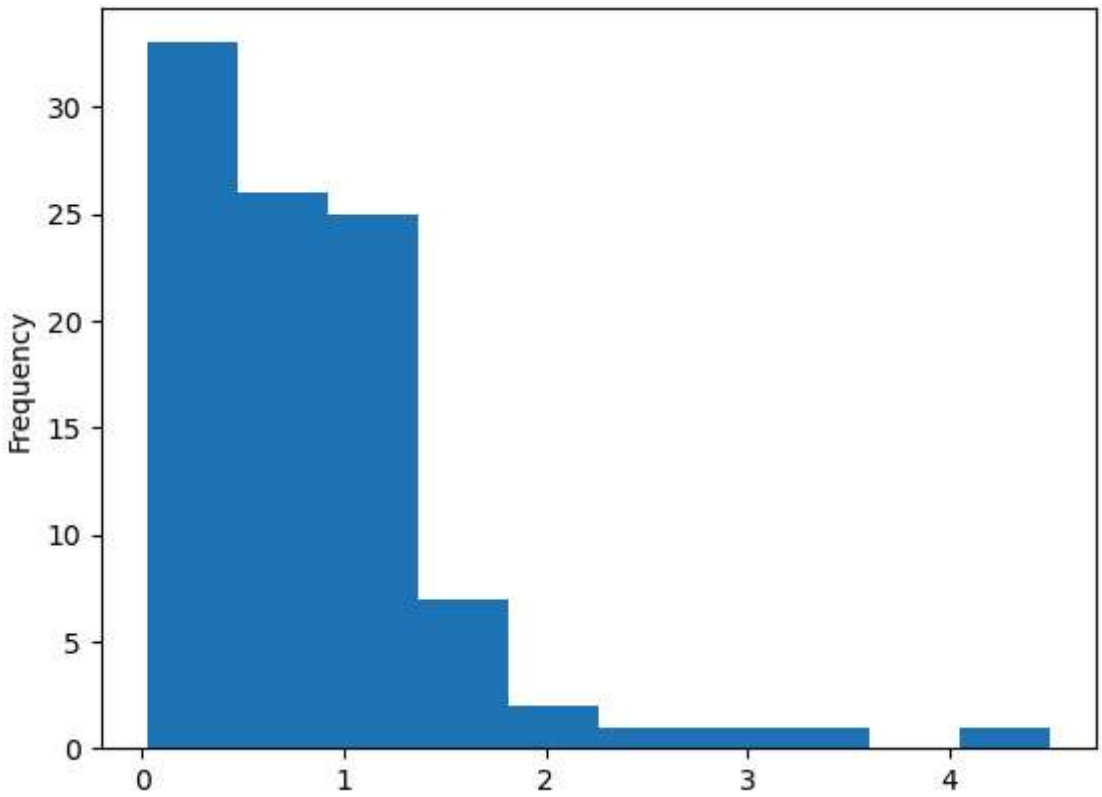
In [140…
```
error_ratio = error_ratio[np.isfinite(error_ratio)]
```

In [141…
```
error_ratio
```

Out[141]:
```
team
AFG    1.000000
ALG    1.000000
ARG    0.853659
ARM    0.428571
AUS    0.367347
         ...
UKR    0.951220
USA    0.126953
UZB    0.625000
VEN    1.750000
VIE    1.000000
Length: 97, dtype: float64
```

In [142…
```
error_ratio.plot.hist()
```

Out[142]:
```
<Axes: ylabel='Frequency'>
```



In [143…
```
error_ratio.sort_values()
```

Out[143]:
```
team
FRA    0.022472
CAN    0.048387
NZL    0.063492
RUS    0.082353
ITA    0.121429
         ...
MAR    2.000000
EGY    2.400000
HKG    3.000000
POR    3.333333
AUT    4.500000
Length: 97, dtype: float64
```

In [ ]: