In [73]:
```
import pandas as pd
```

In [74]:
```
df = pd.read_csv('premier-league-data.csv')
```

In [75]:
```
df.head()
```

Out[75]:

	home_team	away_team	home_goals	away_goals	result	season
0	Sheffield United	Liverpool	1	1	D	2006-2007
1	Arsenal	Aston Villa	1	1	D	2006-2007
2	Everton	Watford	2	1	H	?
3	Newcastle United	Wigan Athletic	2	1	H	2006-2007
4	Portsmouth	Blackburn Rovers	3	0	H	2006-2007

Data Cleaning

Remove invalid values from the `season` column

In [76]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4560 entries, 0 to 4559
Data columns (total 6 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   home_team   4560 non-null   object
 1   away_team   4560 non-null   object
 2   home_goals  4560 non-null   int64 
 3   away_goals  4560 non-null   int64 
 4   result      4560 non-null   object
 5   season      4560 non-null   object
dtypes: int64(2), object(4)
memory usage: 213.9+ KB
```

In [77]:
```
df['season'].value_counts()
```

Out[77]:
```
season
2007-2008    380
2008-2009    380
2009-2010    380
2010-2011    380
2011-2012    380
2012-2013    380
2013-2014    380
2014-2015    380
2015-2016    380
2016-2017    380
2017-2018    380
2006-2007    349
?             31
Name: count, dtype: int64
```

In [78]:
```
filt = df['season'] == '?'
df.loc[filt, 'season'] = 'Unknown season'
```

In [79]:
```
df['season'].value_counts()
```

Out[79]:
```
season
2007-2008         380
2008-2009         380
2009-2010         380
2010-2011         380
2011-2012         380
2012-2013         380
2013-2014         380
2014-2015         380
2015-2016         380
2016-2017         380
2017-2018         380
2006-2007         349
Unknown season     31
Name: count, dtype: int64
```

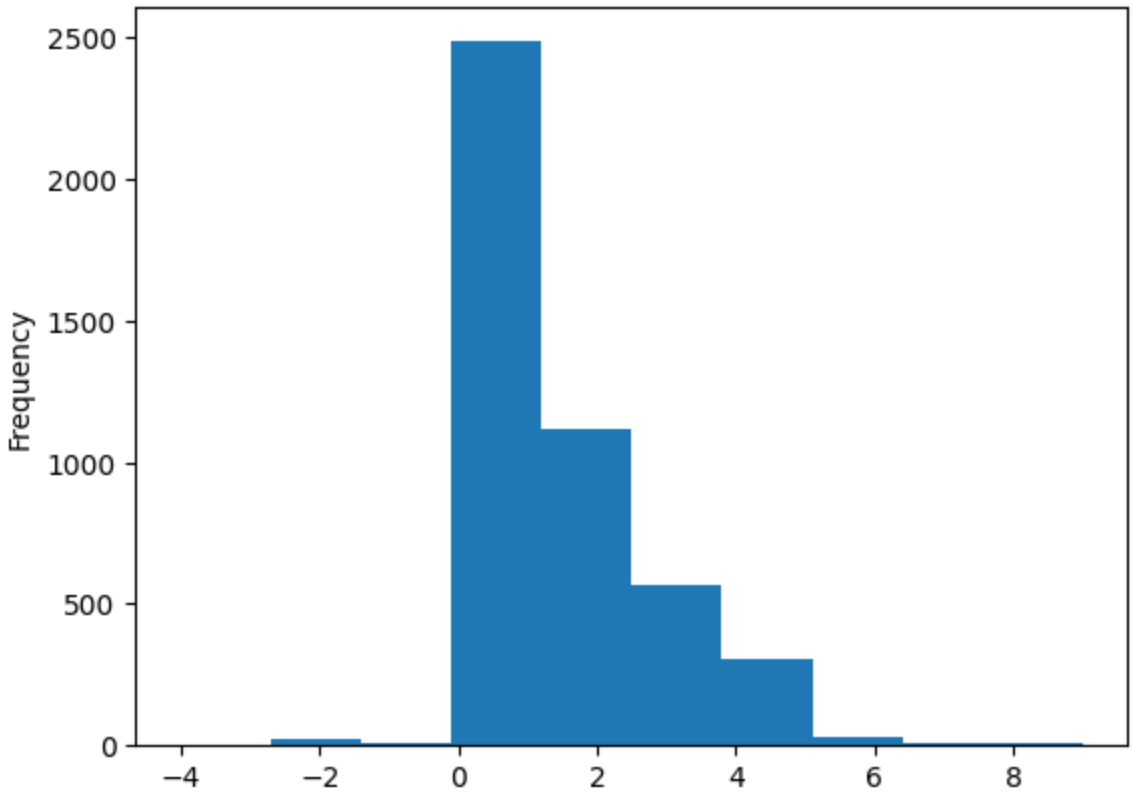
Identify invalid values in goals scored

In [80]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4560 entries, 0 to 4559
Data columns (total 6 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   home_team   4560 non-null   object
 1   away_team   4560 non-null   object
 2   home_goals  4560 non-null   int64 
 3   away_goals  4560 non-null   int64 
 4   result      4560 non-null   object
 5   season      4560 non-null   object
dtypes: int64(2), object(4)
memory usage: 213.9+ KB
```

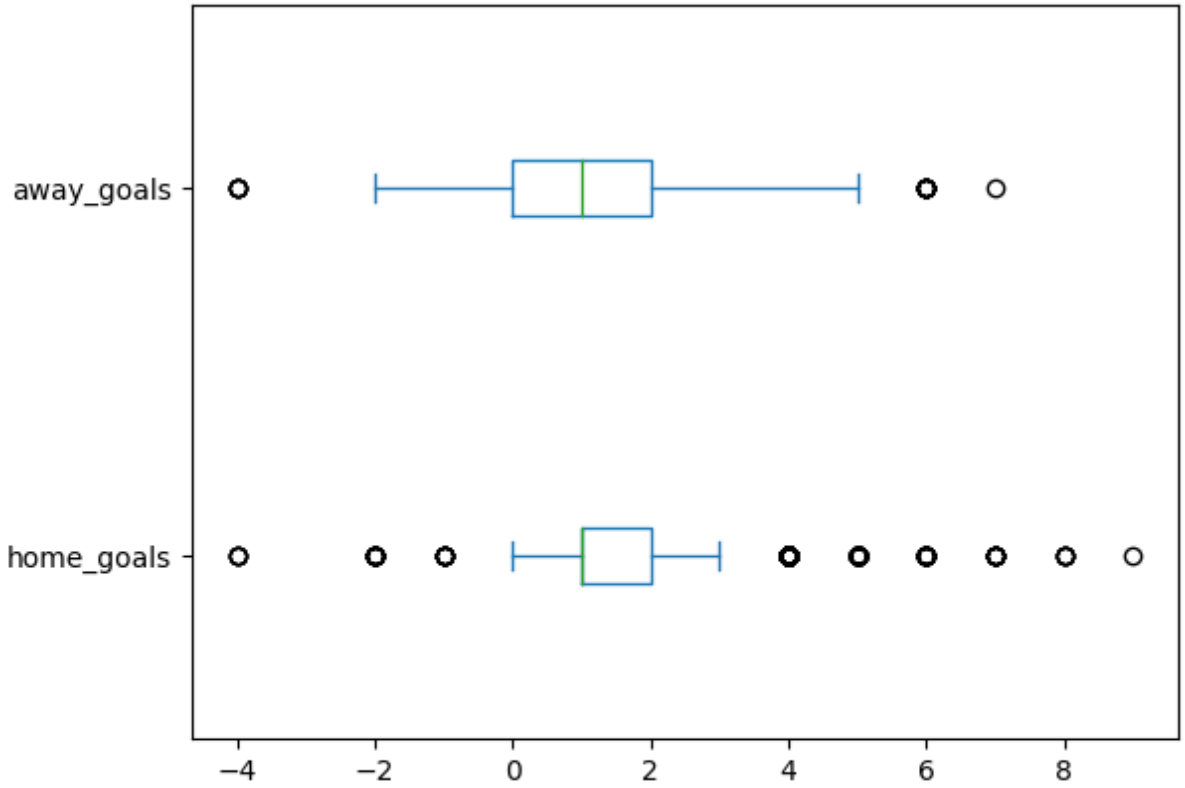
In [81]:
```
df['home_goals'].plot(kind='hist')
```

Out[81]: <Axes: ylabel='Frequency'>



In [82]:
```
df[['home_goals', 'away_goals']].plot(kind='box', vert=False)
```

Out[82]: <Axes: >



In [83]:
```
(df[['home_goals', 'away_goals']] < 0).sum()
```

Out[83]:
```
home_goals    34
away_goals    39
dtype: int64
```

In [84]:
```
filt1 = df['home_goals'] < 0
df.loc[filt1, 'home_goals'].count()
```

Out[84]: 34

In [85]:
```
filt2 = df['away_goals'] < 0
df.loc[filt2, 'away_goals'].count()
```

Out[85]: 39

Replace invalid goals for 0

In [86]:
```
df.loc[filt1, 'home_goals'] = 0
df.loc[filt2, 'away_goals'] = 0
```

Identify and clean invalid results in the `result` column

In [87]:
```
df['result'].value_counts()
```

Out[87]:
```
result
H    2088
A    1278
D    1151
?      43
Name: count, dtype: int64
```

In [88]:
```
filt1 = df['home_goals'] > df['away_goals']
filt2 = df['home_goals'] < df['away_goals']
filt3 = df['home_goals'] == df['away_goals']

df.loc[filt1, 'result'] = 'H'
df.loc[filt2, 'result'] = 'A'
df.loc[filt3, 'result'] = 'D'
```

In [89]:
```
df['result'].value_counts()
```

Out[89]:
```
result
H    2107
A    1294
D    1159
Name: count, dtype: int64
```

Analysis

What's the average number of goals per match?

In [91]:
```
df.head()
```

Out[91]:

	home_team	away_team	home_goals	away_goals	result	season
0	Sheffield United	Liverpool	1	1	D	2006-2007
1	Arsenal	Aston Villa	1	1	D	2006-2007
2	Everton	Watford	2	1	H	Unknown season
3	Newcastle United	Wigan Athletic	2	1	H	2006-2007
4	Portsmouth	Blackburn Rovers	3	0	H	2006-2007

In [92]:
```
(df['home_goals']+df['away_goals']).mean()
```

Out[92]:
```
2.6633771929824563
```

Create a new column `total_goals`

In [93]:
```
df['total_goals'] = df['home_goals']+df['away_goals']
```

In [94]:
```
df.head()
```

Out[94]:

	home_team	away_team	home_goals	away_goals	result	season	total_goals
0	Sheffield United	Liverpool	1	1	D	2006-2007	2
1	Arsenal	Aston Villa	1	1	D	2006-2007	2
2	Everton	Watford	2	1	H	Unknown season	3
3	Newcastle United	Wigan Athletic	2	1	H	2006-2007	3
4	Portsmouth	Blackburn Rovers	3	0	H	2006-2007	3

Calculate average goals per season

In [96]:
```
goals_per_season = df.groupby('season')['total_goals'].mean().sort_index
goals_per_season
```

Out[96]:
```
season
2006-2007         2.429799
2007-2008         2.618421
2008-2009         2.463158
2009-2010         2.747368
2010-2011         2.797368
2011-2012         2.763158
2012-2013         2.773684
2013-2014         2.718421
2014-2015         2.500000
2015-2016         2.676316
2016-2017         2.794737
2017-2018         2.678947
Unknown season    2.419355
Name: total_goals, dtype: float64
```

What's the biggest goal difference in a match?

In [97]:
```
(df['home_goals'] - df['away_goals']).max()
```

Out[97]: 8

In [98]:
```
(df['home_goals'] - df['away_goals']).min()
```

Out[98]: -6

In [99]:
```
(df['home_goals'] - df['away_goals']).abs().max()
```

Out[99]: 8

What's the team with most away wins?

In [112]:
```
filt = df['result'] == 'A'
df[filt].groupby('away_team')['result'].count().sort_values(ascending=Fa
```

Out[112]:
```
away_team
Chelsea                     120
Manchester United           117
Arsenal                     103
Liverpool                    98
Manchester City              98
Tottenham Hotspur            90
Everton                      66
Aston Villa                  53
West Ham United              43
Newcastle United             41
Stoke City                   36
Sunderland                   35
West Bromwich Albion         34
Southampton                  33
Swansea City                 31
Wigan Athletic               29
Crystal Palace               27
Blackburn Rovers             27
Bolton Wanderers             26
Fulham                       23
Leicester City               22
Portsmouth                   16
Watford                      15
AFC Bournemouth              13
Hull City                    13
Burnley                      13
Norwich City                 12
Reading                      10
Birmingham City              10
Wolverhampton Wanderers       9
Middlesbrough                 8
Queens Park Rangers           7
Blackpool                     5
Sheffield United              3
Huddersfield Town             3
Cardiff City                  2
Brighton and Hove Albion      2
Charlton Athletic             1
Name: result, dtype: int64
```

In [116]:
```
df.groupby('away_team').apply(lambda rows: (rows['result'] == 'A')
                              .sum()).sort_values(ascending=False)
```

Out[116]:
```
away_team
Chelsea                     120
Manchester United           117
Arsenal                     103
Manchester City              98
Liverpool                    98
Tottenham Hotspur            90
Everton                      66
Aston Villa                  53
West Ham United              43
Newcastle United             41
Stoke City                   36
Sunderland                   35
West Bromwich Albion         34
Southampton                  33
Swansea City                 31
Wigan Athletic               29
Crystal Palace               27
Blackburn Rovers             27
Bolton Wanderers             26
Fulham                       23
Leicester City               22
Portsmouth                   16
Watford                      15
AFC Bournemouth              13
Hull City                    13
Burnley                      13
Norwich City                 12
Reading                      10
Birmingham City              10
Wolverhampton Wanderers       9
Middlesbrough                 8
Queens Park Rangers           7
Blackpool                     5
Sheffield United              3
Huddersfield Town             3
Cardiff City                  2
Brighton and Hove Albion      2
Charlton Athletic             1
Derby County                  0
dtype: int64
```

What's the team with the most goals scored at home?

In [120]:
```
df.groupby('home_team')['home_goals'].sum().sort_values(ascending=False)
```

Out[120]:
```
home_team
Manchester City             499
Manchester United           495
Chelsea                     488
Arsenal                     471
Liverpool                   459
Tottenham Hotspur           414
Everton                     392
West Ham United             283
Newcastle United            267
Stoke City                  244
Aston Villa                 227
West Bromwich Albion        225
Sunderland                  222
Fulham                      211
Swansea City                179
Southampton                 171
Blackburn Rovers            155
Bolton Wanderers            152
Wigan Athletic              140
Leicester City              119
Crystal Palace              111
Hull City                   107
Portsmouth                  102
Norwich City                 96
Middlesbrough                92
Watford                      91
AFC Bournemouth              84
Burnley                      80
Reading                      71
Birmingham City              67
Wolverhampton Wanderers      62
Queens Park Rangers          60
Blackpool                    30
Brighton and Hove Albion     24
Sheffield United             23
Cardiff City                 20
Charlton Athletic            19
Huddersfield Town            16
Derby County                 12
Name: home_goals, dtype: int64
```

- ***What's the team that received the least amount of goals while playing at home?***

In [134]:
```
new_df = df.groupby('home_team')[['home_team', 'away_goals']].agg({
    'home_team': 'size',
    'away_goals': 'sum'
})
.rename(
    columns={'home_team':'total_games', 'away_goals':'goals_received'})
.sort_values(
    by=['total_games', 'goals_received'], ascending=[False, True])
```

In [135]:
```
new_df
```

Out[135]:

home_team	total_games	goals_received
Manchester United	228	158
Liverpool	228	180
Arsenal	228	183
Chelsea	228	183
Manchester City	228	186
Tottenham Hotspur	228	218
Everton	228	238
West Ham United	209	279
Stoke City	190	213
Aston Villa	190	234

In [139]:
```
new_df['goals_received_per_match'] =
new_df['goals_received'] / new_df['total_games']
```

In [140]:
```
new_df['goals_received_per_match'].sort_values(ascending=True)
```

Out[140]:
```
home_team
Manchester United           0.692982
Liverpool                   0.789474
Arsenal                     0.802632
Chelsea                     0.802632
Manchester City             0.815789
Tottenham Hotspur           0.956140
Birmingham City             1.017544
Everton                     1.043860
Charlton Athletic           1.052632
Sheffield United            1.105263
Burnley                     1.118421
Stoke City                  1.121053
Leicester City              1.131579
Southampton                 1.131579
Blackburn Rovers            1.157895
Portsmouth                  1.184211
Middlesbrough               1.184211
Aston Villa                 1.231579
Sunderland                  1.242105
Newcastle United            1.263158
Fulham                      1.276316
Norwich City                1.276316
Swansea City                1.278195
Huddersfield Town           1.315789
Brighton and Hove Albion    1.315789
Crystal Palace              1.315789
Bolton Wanderers            1.324561
Queens Park Rangers         1.333333
Reading                     1.333333
West Ham United             1.334928
Watford                     1.355263
West Bromwich Albion        1.385965
Wigan Athletic              1.413534
Hull City                   1.494737
AFC Bournemouth             1.561404
Wolverhampton Wanderers     1.666667
Cardiff City                1.736842
Blackpool                   1.947368
Derby County                2.263158
Name: goals_received_per_match, dtype: float64
```

What's the team with most goals scored playing as a visitor (away from home)?

In [142]:
```
df.groupby('away_team')['away_goals']
.sum().sort_values(ascending=False)
```

Out[142]:
```
away_team
Arsenal                     379
Manchester United           366
Manchester City             359
Chelsea                     357
Liverpool                   348
Tottenham Hotspur           339
Everton                     255
Aston Villa                 214
West Ham United             209
Newcastle United            177
Sunderland                  170
West Bromwich Albion        154
Stoke City                  150
Fulham                      127
Swansea City                127
Wigan Athletic              125
Southampton                 123
Blackburn Rovers            122
Bolton Wanderers            111
Crystal Palace              103
Leicester City               98
Hull City                    72
Reading                      65
Burnley                      64
Norwich City                 63
Portsmouth                   63
Watford                      62
AFC Bournemouth              60
Wolverhampton Wanderers      56
Queens Park Rangers          55
Birmingham City              53
Middlesbrough                49
Blackpool                    25
Charlton Athletic            15
Huddersfield Town            12
Cardiff City                 12
Brighton and Hove Albion     10
Sheffield United              8
Derby County                  8
Name: away_goals, dtype: int64
```