In [10]:
```
import pandas as pd
```

In [11]:
```
df = pd.read_csv('premier-league-data.csv')
```

In [12]:
```
df.head()
```

Out[12]:

	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

In [13]:
```
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
```

Data Cleaning

Remove invalid values from the `season` column

In [14]:
```
df['season'].value_counts()
```

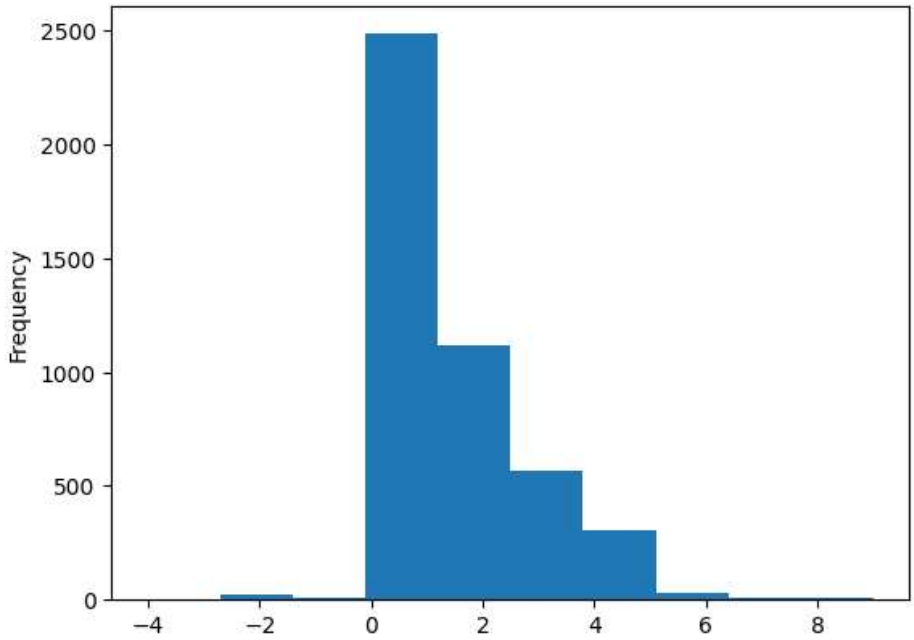
Out[14]:
```
season
2007-2008    380
2008-2009    380
2009-2010    380
2010-2011    380
2011-2012    380
2012-2013    380
2013-2014    380
2014-2015    380
2015-2016    380
2016-2017    380
2017-2018    380
2006-2007    349
?             31
Name: count, dtype: int64
```

In [15]:
```
df.loc[df['season']=='?','season'] = "Unknown season"
```

Identify invalid values in goals scored

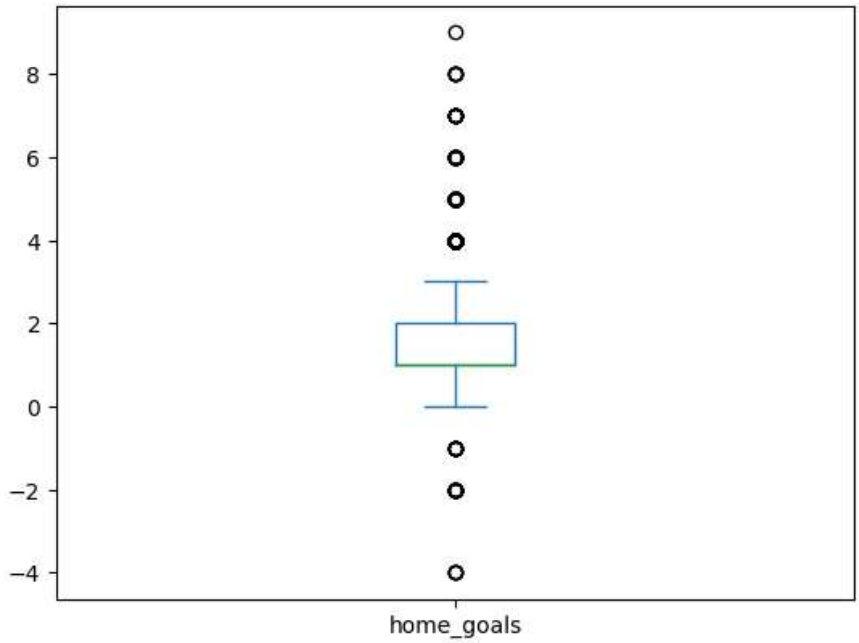
In [17]:
```
df['home_goals'].plot(kind='hist')
```

Out[17]:
```
<Axes: ylabel='Frequency'>
```



In [20]:
```
df['home_goals'].plot(kind='box')
```

Out[20]:
```
<Axes: >
```

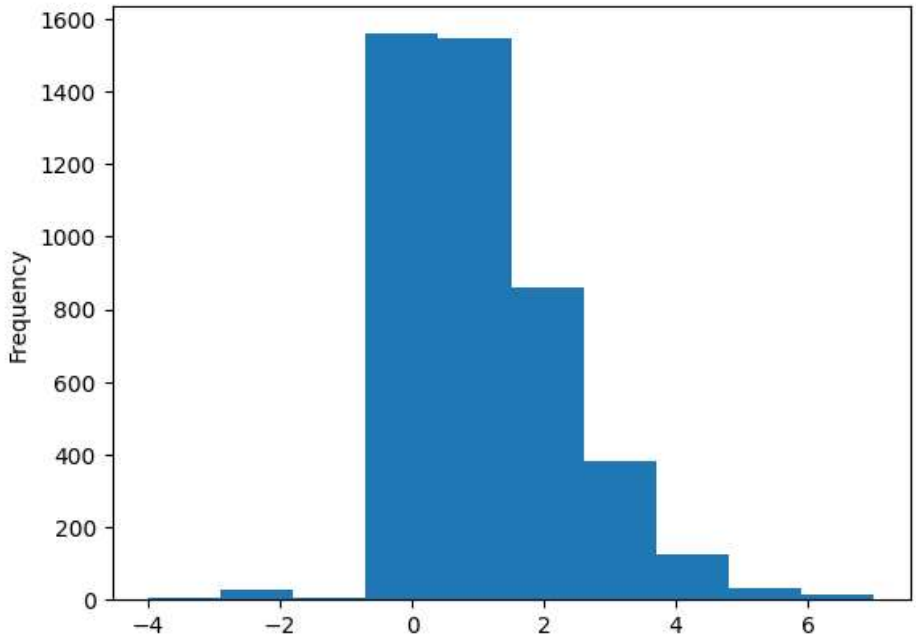


In [21]:
```
df['home_goals'].describe()
```

Out[21]:
```
count    4560.000000
mean        1.516009
std         1.345936
min        -4.000000
25%         1.000000
50%         1.000000
75%         2.000000
max         9.000000
Name: home_goals, dtype: float64
```

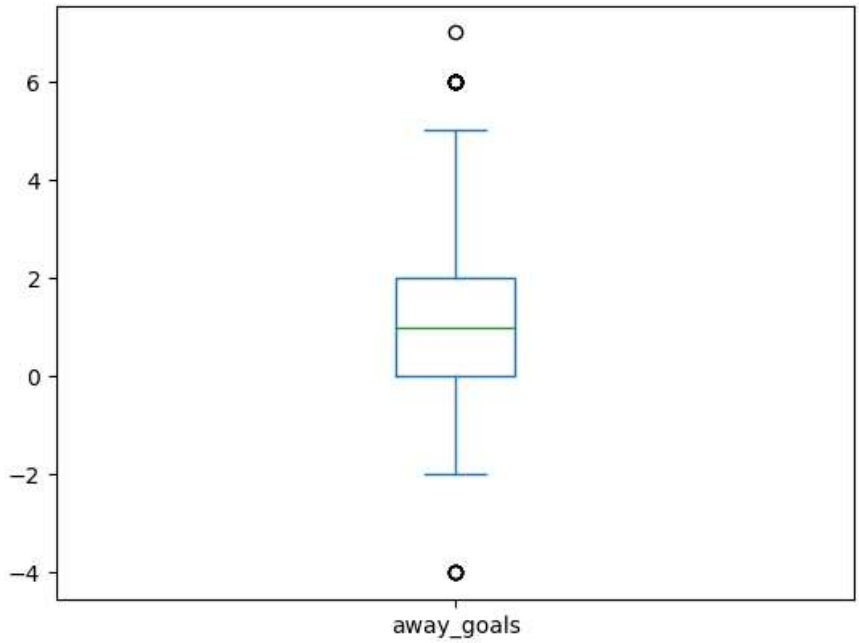
In [18]:
```
df['away_goals'].plot(kind='hist')
```

Out[18]:
```
<Axes: ylabel='Frequency'>
```



In [19]:
```
df['away_goals'].plot(kind='box')
```

Out[19]:
```
<Axes: >
```

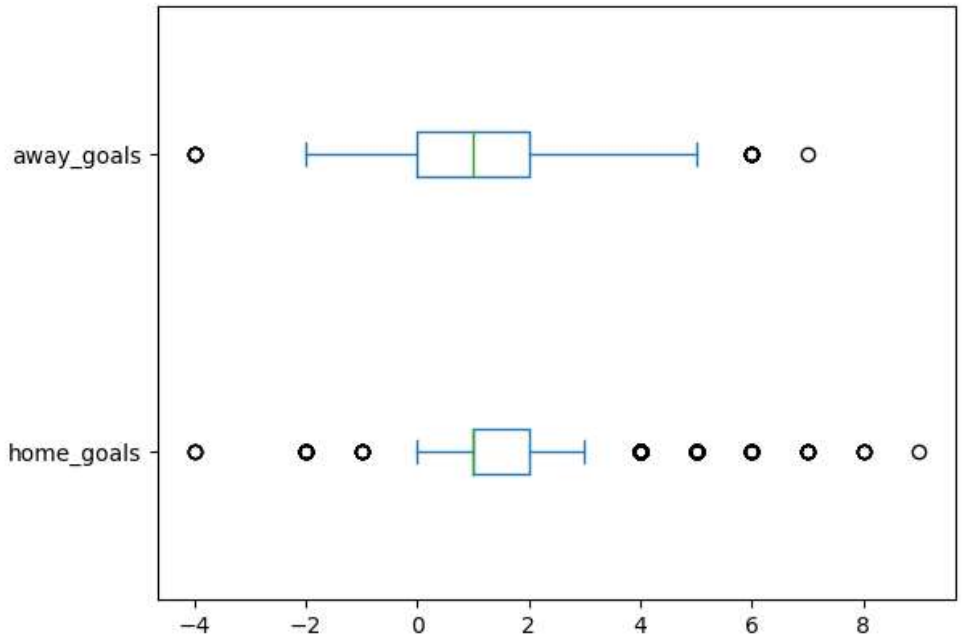


In [22]:
```
df['away_goals'].describe()
```

Out[22]:
```
count    4560.000000
mean        1.114693
std         1.175256
min        -4.000000
25%         0.000000
50%         1.000000
75%         2.000000
max         7.000000
Name: away_goals, dtype: float64
```

In [23]:
```
df[['home_goals','away_goals']].plot(kind='box',vert=False)
```

Out[23]: <Axes: >



In [24]:
```
(df[['home_goals','away_goals']]<0).sum()
```

Out[24]:
```
home_goals    34
away_goals    39
dtype: int64
```

Replace invalid goals for 0

In [25]:
```
df.loc[df['home_goals']<0,'home_goals']=0
df.loc[df['away_goals']<0,'away_goals']=0
```

In [27]:
```
df['season'].value_counts()
```

Out[27]:
```
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 and clean invalid results in the `result` column

In [28]:
```
df['result'].value_counts()
```

Out[28]:
```
result
H    2088
A    1278
D    1151
?      43
Name: count, dtype: int64
```

In [31]:
```
df.loc[df['home_goals']>df['away_goals'],'result']='H'
df.loc[df['home_goals']<df['away_goals'],'result']='A'
df.loc[df['home_goals']==df['away_goals'],'result']='D'
```

In [32]:
```
df['result'].value_counts()
```

Out[32]:
```
result
H    2107
A    1294
D    1159
Name: count, dtype: int64
```

Analysis

What's the average number of goals per match?

In [35]:
```
(df['home_goals']+df['away_goals']).mean()
```

Out[35]:
```
2.6633771929824563
```

Create a new column `total_goals`

In [36]:
```
df['total_goals']=df['home_goals']+df['away_goals']
```

In [37]:
```
df.head()
```

Out[37]:

	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 [40]:
```
df.groupby("season")["total_goals"].mean().sort_index()
```

Out[40]:
```
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
```

In [41]:
```
goals_per_season = df.groupby("season")["total_goals"].mean().sort_index()
```

In [43]:
```
goals_per_season
```

Out[43]:
```
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 [44]:
```
(df['home_goals']-df['away_goals']).max()
```

Out[44]: 8

In [45]:
```
(df['away_goals']-df['home_goals']).max()
```

Out[45]: 6

What's the team with most away wins?

In [52]:
```
df.loc[df['result']=='A'].groupby('away_team')['result'].size().sort_values(ascending=False)
```

Out[52]:
```
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 [54]:
```
df.groupby('away_team').apply(lambda rows:(rows['result']=='A').sum()).sort_values(ascending=False).head
```

Out[54]:
```
away_team
Chelsea             120
Manchester United   117
Arsenal             103
Manchester City      98
Liverpool            98
dtype: int64
```

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

In [ ]:
```

```

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

In [ ]:
```

```

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

In [ ]: