In [1]:
```
import pandas as pd
```

In [2]:
```
df = pd.read_csv('premier-league-data.csv')
```

In [3]:
```
df.head()
```

Out[3]:

	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 [4]:
```
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 [8]:
```
df['season'].value_counts()
```

Out[8]:
```
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 [14]:
```
df.loc[df['season']=='?','season']='Unknown season'
```

In [15]:
```
df['season'].value_counts()
```

Out[15]:
```
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 [17]:
```
df['home_goals'].value_counts()
```

Out[17]:
```
home_goals
 1    1436
 2    1119
 0    1050
 3     568
 4     232
 5      75
 6      30
-2      21
 7      10
-1       9
 8       5
-4       4
 9       1
Name: count, dtype: int64
```

In [18]:
```
df['away_goals'].value_counts()
```

Out[18]:
```
away_goals
 0    1558
 1    1548
 2     862
 3     380
 4     127
 5      32
-2      28
 6      13
-1       6
-4       5
 7       1
Name: count, dtype: int64
```

Replace invalid goals for 0

In [20]:
```
df.loc[df['home_goals']<0,'home_goals']=0
```

In [23]:
```
df.loc[df['away_goals']<0,'away_goals']=0
```

Identify and clean invalid results in the `result` column

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

Out[27]:
```
result
H    2088
A    1278
D    1151
?      43
Name: count, dtype: int64
```

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

In [32]:
```
df.loc[df['home_goals']<df['away_goals'],'result']='A'
```

In [34]:
```
df.loc[df['home_goals']==df['away_goals'],'result']='D'
```

In [35]:
```
df['result'].value_counts()
```

Out[35]:
```
result
H    2107
A    1294
D    1159
Name: count, dtype: int64
```

Analysis

What's the average number of goals per match?

In [42]:
```
(df['home_goals'].sum() + df['away_goals'].sum())/len(df)
```

Out[42]: 2.6633771929824563

In [41]:
```
len(df)
```

Out[41]: 4560

Create a new column `total_goals`

In [43]:
```
df['total_goals'] = df['home_goals'] + df['away_goals']
```

In [44]:
```
df.head()
```

Out[44]:

	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 [50]:
```
df.groupby('season')['total_goals'].mean()
```

Out[50]:
```
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 [51]:
```
goals_per_season = df.groupby('season')['total_goals'].mean()
```

What's the biggest goal difference in a match?

In [53]:
```
df['different_goals'] = abs(df['home_goals']-df['away_goals'])
```

In [54]:
```
df['different_goals'].max()
```

Out[54]: 8

What's the team with most away wins?

In [67]:
```
df.loc[df['result']=='A','away_team'].value_counts().sort_values(ascending=False)
```

Out[67]:

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
Hull City	13
AFC Bournemouth	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: count, dtype: int64
```

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

In [70]:
```
df.groupby('home_team')['home_goals'].sum().sort_values(ascending=False)
```

Out[70]:

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 [80]:
```
s1 = df.groupby('home_team')['away_goals'].sum().sort_values(ascending=True)
```

In [81]:
```
s2 = df['home_team'].value_counts()
```

In [84]:
```
df1 = pd.concat([s1,s2],axis=1)
```

In [85]:
```
df1.head()
```

Out[85]:

home_team	away_goals	count
Charlton Athletic	20	19
Sheffield United	21	19
Brighton and Hove Albion	25	19
Huddersfield Town	25	19
Cardiff City	33	19

In [87]:
```
df1['ratio'] = df1['away_goals']/df1['count']
df1.head()
```

Out[87]:

home_team	away_goals	count	ratio
Charlton Athletic	20	19	1.052632
Sheffield United	21	19	1.105263
Brighton and Hove Albion	25	19	1.315789
Huddersfield Town	25	19	1.315789
Cardiff City	33	19	1.736842

In [91]:
```
df1.loc[df1['ratio']==df1['ratio'].min()]
```

Out[91]:

home_team	away_goals	count	ratio
Manchester United	158	228	0.692982

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

In [95]:
```
df.groupby('away_team')['away_goals'].sum().sort_values(ascending=False)
```

Out[95]:

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
```

In [ ]: