In [1]:

```
import pandas as pd
```

In [17]:

```
df = pd.read_csv('premier-league-data.csv')
```

In [18]:

```
df.head()
```

Out[18]:

	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 [19]:

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

In [20]:

```
df.head()
```

Out[20]:

	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

Identify invalid values in goals scored

In [21]:

```
df.loc[(df['away_goals'] < 0) ,'away_goals'].value_counts().sum()
```

Out[21]:

```
39
```

In [22]:

```
df.loc[(df['home_goals'] < 0), 'home_goals'].value_counts().sum()
```

Out[22]:

```
34
```

Replace invalid goals for 0

In [23]:

```
df.loc[(df['away_goals'] < 0) ,'away_goals'] = 0
df.loc[(df['home_goals'] < 0), 'home_goals'] = 0
```

Identify and clean invalid results in the `result` column

In [24]:

```
df.loc[df['away_goals'] > df['home_goals'], 'result'] = "A"
```

In [25]:

```
df.loc[df['away_goals'] < df['home_goals'], 'result'] = "H"
```

In [26]:

```
df.loc[df['away_goals'] == df['home_goals'], 'result'] = "D"
```

Analysis

What's the average number of goals per match?

In [72]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4560 entries, 0 to 4559
Data columns (total 7 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
 6   total_goals  4560 non-null   int64 
dtypes: int64(3), object(4)
memory usage: 249.5+ KB
```

In [28]:
```
(df['away_goals'].sum()+df['home_goals'].sum())/4560

#4560 is total no. of games
```

Out[28]:
```
2.6633771929824563
```

Create a new column `total_goals`

In [29]:
```
df['total_goals'] = df['away_goals'] + df['home_goals']
```

In [30]:
```
df.head()
```

Out[30]:

	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 [31]:
```
# groupby use krna seekhenge yahan...
df.groupby('season')['total_goals'].mean().sort_index()
```

Out[31]:
```
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 [32]:
```
goals_per_season = df.groupby('season')['total_goals'].mean().sort_index()
```

What's the biggest goal difference in a match?

In [33]:
```
(df['away_goals']-df['home_goals']).max()
```

Out[33]:
```
6
```

In [34]:
```
df.head()
```

Out[34]:

	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

```
In [35]: df.loc[2072]
```

```
Out[35]: home_team      Manchester United
         away_team         Wigan Athletic
         home_goals                     5
         away_goals                     0
         result                         H
         season                 2011-2012
         total_goals                    5
         Name: 2072, dtype: object
```

```
In [36]: (df['home_goals'] - df['away_goals']).sort_values(ascending = False)
```

```
Out[36]: 1514    8
         2458    8
         3116    8
         1265    8
         1497    7
                ..
         712    -6
         3678   -6
         4173   -6
         4225   -6
         2622   -6
         Length: 4560, dtype: int64
```

What's the team with most away wins?

```
In [37]: df.loc[df['result'] == 'A','away_team'].value_counts().head()
```

```
Out[37]: away_team
         Chelsea              120
         Manchester United    117
         Arsenal              103
         Liverpool             98
         Manchester City       98
         Name: count, dtype: int64
```

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

```
In [38]: df.groupby('home_team')['home_goals'].sum().sort_values(ascending = False).head()
```

```
Out[38]: home_team
         Manchester City      499
         Manchester United    495
         Chelsea              488
         Arsenal              471
         Liverpool            459
         Name: home_goals, dtype: int64
```

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

```
In [44]: df.groupby('home_team')['away_goals'].value_counts()
```

```
Out[44]: home_team                away_goals
         AFC Bournemouth          2             15
                                  1             15
                                  0             14
                                  3              9
                                  4              3
                                                ..
         Wolverhampton Wanderers  2             15
                                  3             11
                                  0             10
                                  5              2
                                  4              1
         Name: count, Length: 223, dtype: int64
```

```
In [56]: final = df.groupby('home_team')[['home_team','away_goals']].agg({
             'home_team' : 'size' , 'away_goals' : 'sum'}
             ).rename(columns ={'home_team' : 'total_matches'}
                 ).sort_values(by='total_matches' , ascending = False)
         final.head()
```

Out[56]:

	total_matches	away_goals
home_team		
Liverpool	228	180
Tottenham Hotspur	228	218
Manchester United	228	158
Manchester City	228	186
Arsenal	228	183

In [53]:
```
final['ratio'] = final['away_goals']/ final['total_matches']
```

In [55]:
```
final.sort_values(by = 'ratio').head()
```

Out[55]:

	total_matches	away_goals	ratio
home_team			
Manchester United	228	158	0.692982
Liverpool	228	180	0.789474
Arsenal	228	183	0.802632
Chelsea	228	183	0.802632
Manchester City	228	186	0.815789

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

In [57]:
```
df.head()
```

Out[57]:

	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

In [63]:
```
afinal = df.groupby('away_team')[['away_team','away_goals']].agg({
    'away_team' : 'size' , 'away_goals' : 'sum'}
    ).rename(columns = {'away_team' : 'matches'}).sort_values(
    by = 'matches' , ascending = False)
afinal.head()
```

Out[63]:

	matches	away_goals
away_team		
Liverpool	228	348
Tottenham Hotspur	228	339
Manchester United	228	366
Manchester City	228	359
Arsenal	228	379

In [64]:
```
afinal['ratio'] = afinal['away_goals']/afinal['matches']
afinal.head()
```

Out[64]:

	matches	away_goals	ratio
away_team			
Liverpool	228	348	1.526316
Tottenham Hotspur	228	339	1.486842
Manchester United	228	366	1.605263
Manchester City	228	359	1.574561
Arsenal	228	379	1.662281

In [66]:
```
afinal.sort_values(by = 'ratio',ascending = False).head(3)
```

Out[66]:

	matches	away_goals	ratio
away_team			
Arsenal	228	379	1.662281
Manchester United	228	366	1.605263
Manchester City	228	359	1.574561