LaLiga -Applied Statistics

Domain: Sports Analytics

Context:

La Liga is the men's top professional football division of the Spanish football league system. The dataset contains information on all the teams that have participated in all the past tournaments. It has data about how many goals each team scored, conceded, how many times they came within the first 6 positions, how many seasons they have qualified, their best position in the past, etc.

Objective:

Using Python functions and we want to come up with metrics which can be used to gauge the winning team in the upcoming La Liga cup (Football tournament). Also we want to analyze a few patterns like which team has been most consistent across seasons. Which team has the highest number of goal difference. Which team has the best ranking.

Steps and tasks:

1. Read the data set and replace dashes with 0 to make sure you can perform arithmetic operations on the data. And check the distribution for the 'Best Position' and report the top position (7 points)

In [1]:

```
# importing libraries
import numpy as np
import pandas as pd
```

In [2]:

```
# reading data set from provided CSV to a dataframe 'laliga' using pandas
laliga = pd.read_csv('Laliga_scores.csv')
```

In [3]:

```
# printing head of the dataframe 'laliga'
laliga.head()
```

Out[3]:

	Pos	Team	Seasons	Points	GamesPlayed	GamesWon	GamesDrawn	GamesLost	GoalsFor	Goa
0	1	Real Madrid	86	4385	2762	1647	552	563	5947	
1	2	Barcelona	86	4262	2762	1581	573	608	5900	
2	3	Atletico Madrid	80	3442	2614	1241	598	775	4534	
3	4	Valencia	82	3386	2664	1187	616	861	4398	

	Pos	Team	Seasons	Points	GamesPlayed	GamesWon	GamesDrawn	GamesLost	GoalsFor	Goa
4	5	Athletic Bilbao	86	3368	2762	1209	633	920	4631	

In [4]:

```
#checking shape of the 'laliga' dataframe
laliga.shape
```

Out[4]:

```
(61, 18)
```

In [5]:

```
#checking data types of all columns
laliga.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 61 entries, 0 to 60
Data columns (total 18 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   Pos           61 non-null     int64 
 1   Team          61 non-null     object
 2   Seasons       61 non-null     int64 
 3   Points        61 non-null     object
 4   GamesPlayed   61 non-null     object
 5   GamesWon      61 non-null     object
 6   GamesDrawn    61 non-null     object
 7   GamesLost     61 non-null     object
 8   GoalsFor      61 non-null     object
 9   GoalsAgainst  61 non-null     object
 10  Champion      61 non-null     object
 11  Runner-up     61 non-null     object
 12  Third         61 non-null     object
 13  Fourth        61 non-null     object
 14  Fifth         61 non-null     object
 15  Sixth         61 non-null     object
 16  Debut         61 non-null     object
 17  BestPosition  61 non-null     int64 
dtypes: int64(3), object(15)
memory usage: 8.7+ KB
```

In [6]:

```
# using replace function to replace '-' with '0' which will allow us arithamatic operat
laliga.replace('-',0,inplace=True)
laliga
```

Out[6]:

	Pos	Team	Seasons	Points	GamesPlayed	GamesWon	GamesDrawn	GamesLost	GoalsFor	Go
0	1	Real Madrid	86	4385	2762	1647	552	563	5947	
1	2	Barcelona	86	4262	2762	1581	573	608	5900	
2	3	Atletico Madrid	80	3442	2614	1241	598	775	4534	
3	4	Valencia	82	3386	2664	1187	616	861	4398	
4	5	Athletic Bilbao	86	3368	2762	1209	633	920	4631	

	Pos	Team	Seasons	Points	GamesPlayed	GamesWon	GamesDrawn	GamesLost	GoalsFor	Go
...	...	...	...	...	...	...	...	...	...	
56	57	Xerez	1	34	38	8	10	20	38	
57	58	Condal	1	22	30	7	8	15	37	
58	59	Atletico Tetuan	1	19	30	7	5	18	51	
59	60	Cultural Leonesa	1	14	30	5	4	21	34	
60	61	Girona	1	0	0	0	0	0	0	

61 rows × 18 columns

2. Print all the teams which have started playing between 1930-1980. Use “Debut” column. (Include year 1930 only)

In [7]:
```
# converting values of 'Debut' column into string datatype
laliga['Debut'] = laliga['Debut'].astype(str)
```

In [8]:
```
# getting team details to new dataframe 'Debut Year' based on Debut in between 1930 to
Debut_Year = laliga[laliga['Debut'].str[:4].between('1930','1980')]
```

In [31]:
```
# printing team name and debut year from 'Debut_Year' dataframe
Debut_Year[['Team','Debut']]
```

Out[31]:

	Team	Debut
3	Valencia	1931-32
5	Sevilla	1934-35
8	Zaragoza	1939-40
9	Real Betis	1932-33
10	Deportivo La Coruna	1941-42
11	Celta Vigo	1939-40
12	Valladolid	1948-49
14	Sporting Gijon	1944-45
15	Osasuna	1935-36
16	Malaga	1949-50
17	Oviedo	1933-34
18	Mallorca	1960-61
19	Las Palmas	1951-52

	Team	Debut
21	Granada	1941-42
22	Rayo Vallecano	1977-78
23	Elche	1959-60
25	Hercules	1935-36
26	Tenerife	1961-62
27	Murcia	1940-41
28	Alaves	1930-31
29	Levante	1963-64
30	Salamanca	1974-75
31	Sabadell	1943-44
32	Cadiz	1977-78
34	Castellon	1941-42
37	Cordoba	1962-63
39	Recreativo	1978-79
40	Burgos CF	1971-72
41	Pontevedra	1963-64
46	Gimnastic	1947-48
49	Alcoyano	1945-46
50	Jaen	1953-54
52	AD Almeria	1979-80
54	Lleida	1950-51
57	Condal	1956-57
58	Atletico Tetuan	1951-52
59	Cultural Leonesa	1955-56

3. Print the list of teams which came Top 5 in terms of points (5 points)

In [10]:

```
# copying 'Team' and 'Points' column to new dataframe 'laliga_sort'
laliga_sort = laliga[['Team','Points']].copy()
```

In [11]:

```
# converting values of 'Points' column into int datatype
laliga_sort['Points'] = laliga_sort['Points'].astype(int)
```

In [12]:

```
# sorting the dataframe 'laliga_sort' based on 'Points' value
laliga_sort.sort_values(by='Points', ascending=False, inplace=True)
```

In [13]:

```
#printing head as top 5 teams in terms of points
```

```
laliga_sort.head(5)
```

Out[13]:

```
             Team  Points
0     Real Madrid    4385
1       Barcelona    4262
2 Atletico Madrid    3442
3        Valencia    3386
4 Athletic Bilbao    3368
```

4. Write a function with the name “Goal_diff_count” which should return all the teams with their Goal Differences. (5 points)

In [14]:

```
# converting values of 'GoalsFor' and 'GoalsAgainst' column into int datatype
laliga['GoalsFor'] = laliga['GoalsFor'].astype(int)
laliga['GoalsAgainst'] = laliga['GoalsAgainst'].astype(int)
```

In [15]:

```
# defining function 'Goal_diff_count()' as per the problem statement needs
def Goal_diff_count():
    laliga['Goal_diff_count'] = laliga['GoalsFor']-laliga['GoalsAgainst']
    return laliga[['Team','Goal_diff_count']]
```

In [16]:

```
# calling 'Goal_diff_count()' function and storing it to 'Goal'
Goal = Goal_diff_count()
```

In [17]:

```
# sorting data and printing data from 'Goal'
Goal.sort_values(by = 'Goal_diff_count',ascending=False)
```

Out[17]:

```
               Team  Goal_diff_count
 0      Real Madrid             2807
 1        Barcelona             2786
 2  Atletico Madrid             1225
 4  Athletic Bilbao              931
 3         Valencia              929
...             ...              ...
27           Murcia             -385
19       Las Palmas             -399
14   Sporting Gijon             -399
12        Valladolid            -413
13 Racing Santander             -525
```

61 rows × 2 columns

In [18]:

```
#new column 'Goal_diff_count' added to dataframe so shape is changed from (61, 20)to (6
laliga.shape
```

Out[18]:

```
(61, 19)
```

In [19]:

```
# conforming addition of new column to dataframe by printing head
laliga.head()
```

Out[19]:

	Pos	Team	Seasons	Points	GamesPlayed	GamesWon	GamesDrawn	GamesLost	GoalsFor	Goa
0	1	Real Madrid	86	4385	2762	1647	552	563	5947	
1	2	Barcelona	86	4262	2762	1581	573	608	5900	
2	3	Atletico Madrid	80	3442	2614	1241	598	775	4534	
3	4	Valencia	82	3386	2664	1187	616	861	4398	
4	5	Athletic Bilbao	86	3368	2762	1209	633	920	4631	

5. Using the same function, find the team which has a maximum and minimum goal difference. (5 points)

In [20]:

```
# printed first entry of sorted dataframe by calling 'Goal_diff_count()' function
Goal_diff_count().head(1) # max goal difference
```

Out[20]:

	Team	Goal_diff_count
0	Real Madrid	2807

In [21]:

```
# printed last entry of sorted dataframe by calling 'Goal_diff_count()' function
Goal_diff_count().tail(1) # min goal difference
```

Out[21]:

	Team	Goal_diff_count
60	Girona	0

6. Create a new column with the name “Winning Percent” and append it to the data set (7 points) Percentage of Winning = (GamesWon / GamesPlayed)*100

If there are any numerical error, replace it with 0%

In [22]:

```
# converting values of 'GamesWon' and 'GamesPlayed' column into int datatype
laliga['GamesWon'] = laliga['GamesWon'].astype(int)
laliga['GamesPlayed'] = laliga['GamesPlayed'].astype(int)
```

In [23]:

```
# calculate winning percentage for each team and storing value to new column 'Win Per'
laliga['Win Per'] = (laliga['GamesWon']/laliga['GamesPlayed']) *100
```

In [24]:

```
# replacing Null values with 0%
laliga['Win Per'].fillna(0,inplace = True)
```

In [25]:

```
# printing team name and winning percentage for each team
laliga[['Team','Win Per']]
```

Out[25]:

	Team	Win Per
0	Real Madrid	59.630702
1	Barcelona	57.241130
2	Atletico Madrid	47.475134
3	Valencia	44.557057
4	Athletic Bilbao	43.772629
...	...	...
56	Xerez	21.052632
57	Condal	23.333333
58	Atletico Tetuan	23.333333
59	Cultural Leonesa	16.666667
60	Girona	0.000000

61 rows × 2 columns

7. Print the top 5 teams which have the highest Winning percentage (5 points)

In [26]:

```
# printing top 5 teams with highest winning percentage
laliga[['Team','Win Per']].head(5)
```

Out[26]:

	Team	Win Per
0	Real Madrid	59.630702
1	Barcelona	57.241130
2	Atletico Madrid	47.475134
3	Valencia	44.557057
4	Athletic Bilbao	43.772629

8. Group teams based on their “Best position” and print the sum of their points for all positions.

In [27]:

```
# converting values of 'Points' and 'BestPosition' column into int datatype
laliga['Points'] = laliga['Points'].astype(int)
laliga['BestPosition'] = laliga['BestPosition'].astype(int)
```

In [28]:

```
# grouping teams based on 'BeastPosition' column
group_Best = laliga[['Team','Points','BestPosition']].groupby('BestPosition')
```

In [29]:

```
# computing sum of grouped values on 'BestPosition' and print them
group_Best['Points'].sum()
```

```
Out[29]: BestPosition
         1     27933
         2      6904
         3      5221
         4      6563
         5      1884
         6      2113
         7      1186
         8      1134
         9        96
         10      450
         11      445
         12      511
         14       71
         15       14
         16       81
         17      266
         19       81
         20       34
         Name: Points, dtype: int32
```