Applied Statistics on LaLiga Dataset

Domain:-Sports

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.

Data Description:- Laliga.csv => The data set contains information on all the teams so far participated in all the past tournaments

Objective:- We want to use statistical techniques to come up with metrics with which can be used to gauge the winning team in the upcoming La Liga cup (Football tournament)

In [1]:
```
#Importing Necessary Libraries
import pandas as pd
import numpy as np
```

Task 1: **Read the data set and replace dashes with 0 to make sure you can perform arithmetic operations on the data. (10 points)**

In [2]:
```
#Read the data from file with name "Laliga.csv" using the pandas library to print data in Dataframe.
laliga = pd.read_csv("Laliga.csv")
```

In [3]:
```
#Print First 5 Rows of DataFrame
laliga.head()
```

Out[3]:

	Pos	Team	Seasons	Points	GamesPlayed	GamesWon	GamesDrawn	GamesLost	GoalsFor	GoalsAgainst	Champion	Ru
0	1	Real Madrid	86	4385	2762	1647	552	563	5947	3140	33	
1	2	Barcelona	86	4262	2762	1581	573	608	5900	3114	25	
2	3	Atletico Madrid	80	3442	2614	1241	598	775	4534	3309	10	
3	4	Valencia	82	3386	2664	1187	616	861	4398	3469	6	
4	5	Athletic Bilbao	86	3368	2762	1209	633	920	4631	3700	8	

In [4]:
```
#Check for Presence of any Null Values in DataFrame
laliga.isnull().sum()
```

Out[4]:
```
Pos             0
Team            0
Seasons         0
Points          0
GamesPlayed     0
GamesWon        0
GamesDrawn      0
GamesLost       0
GoalsFor        0
GoalsAgainst    0
Champion        0
Runner-up       0
Third           0
Fourth          0
Fifth           0
Sixth           0
T               0
Debut           0
Since/LastApp   0
BestPosition    0
dtype: int64
```

In [5]:
```
#Print a concise summary of a DataFrame
laliga.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 61 entries, 0 to 60
Data columns (total 20 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  T              61 non-null     object
 17  Debut          61 non-null     object
 18  Since/LastApp  61 non-null     object
 19  BestPosition   61 non-null     int64
dtypes: int64(3), object(17)
memory usage: 9.7+ KB
```

In [6]:
```
#Replacing Hyphens('-') present in Dataset with Value 0, so that we can perform arithmetic operations on the da
laliga.replace('-',0,inplace=True)
laliga.tail()
```

Out[6]:

	Pos	Team	Seasons	Points	GamesPlayed	GamesWon	GamesDrawn	GamesLost	GoalsFor	GoalsAgainst	Champion	Ru
56	57	Xerez	1	34	38	8	10	20	38	66	0	
57	58	Condal	1	22	30	7	8	15	37	57	0	
58	59	Atletico Tetuan	1	19	30	7	5	18	51	85	0	
59	60	Cultural Leonesa	1	14	30	5	4	21	34	65	0	
60	61	Girona	1	0	0	0	0	0	0	0	0	

In [7]:
```
#Check the Number of Rows and Columns present in DataFrame.
laliga.shape
```

Out[7]: (61, 20)

Task 2: **Print all the teams which have started playing between 1930-1980. Use “Debut” column (Include year 1930 only) (10 points)**

In [8]:
```
#Converting Type of Values present in Column "Debut" to String.
laliga['Debut'] = laliga['Debut'].astype(str)

#Storing Teams with Debut Year in between 1930-1980 (Note: Year 1930 is included) in a new DataFrame
Debut_Year = laliga[laliga['Debut'].str[:4].between('1930','1980')]

#Print Team Name and its Debut Year which is in between 1930-1980 (Note: Year 1930 is included).
Debut_Year[['Team','Debut']]
```

Out[8]:

	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
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

Task 3: **Print the list of teams which came Top 5 in terms of points (5 points)**

In [9]:
```
#Creating a new dataframe containing Teams with their Points
df_points = laliga[['Team','Points']].copy()

#Converting Type of Values present in Column "Points" to Integer.
df_points['Points'] = df_points['Points'].astype(int)

#Sorting the Dataframe "df_points" on values of Points
df_points.sort_values(by='Points', ascending=False, inplace=True)

#Print the list of top 5 teams which have the highest Points.
df_points.head(5)
```

Out[9]:

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

Task 4: **Write a function with the name "Goal_diff_count" which should return all the teams with their Goal Differences. (5 points)**

In [10]:
```
#Converting Type of Values present in Column "GoalsFor" to Integer.
laliga['GoalsFor'] = laliga['GoalsFor'].astype(int)

#Converting Type of Values present in Column "GoalsAgainst" to Integer.
laliga['GoalsAgainst'] = laliga['GoalsAgainst'].astype(int)

#Create a function with the name "Goal_diff_count" which return Teams with their Goal Difference
def Goal_diff_count():
     laliga['Goal_Difference'] = laliga['GoalsFor']-laliga['GoalsAgainst']
     return laliga[['Team','Goal_Difference']]

#Call the Function "Goal_diff_count()"
Goal = Goal_diff_count()
Goal
```

Out[10]:

	Team	Goal_Difference
0	Real Madrid	2807
1	Barcelona	2786
2	Atletico Madrid	1225
3	Valencia	929
4	Athletic Bilbao	931
...	...	...
56	Xerez	-28
57	Condal	-20
58	Atletico Tetuan	-34
59	Cultural Leonesa	-31
60	Girona	0

61 rows × 2 columns

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

Hint: Goal_diff_count = GoalsFor - GoalsAgainst

In [11]:
```
#Sorting the Dataframe "Goal" on values of Goal_Difference
df_gd = Goal.sort_values(by = 'Goal_Difference',ascending=False)

#Storing Team with Maximum Goal Difference in a new variable "df_gd_maximum".
df_gd_maximum = df_gd.head(1)

#Storing Team with Minimum Goal Difference in a new variable "df_gd_minimum".
df_gd_minimum = df_gd.tail(1)

#Print the Teams with Maximum and Minimum Goal Difference.
print("=> Team in Laliga with Maximum Goal Difference is: \n\n",df_gd_maximum)
print("\n=> Team in Laliga with Maximum Goal Difference is:\n\n",df_gd_minimum)
```

```
=> Team in Laliga with Maximum Goal Difference is:

          Team  Goal_Difference
0  Real Madrid             2807

=> Team in Laliga with Maximum Goal Difference is:

                Team  Goal_Difference
13  Racing Santander             -525
```

Task 6: **Create a new column with the name "Winning Percent" and append it to the data set. If there are any numerical error, replace it with 0% (5 points)**

Hint: Percentage of Winning = (GamesWon / GamesPlayed)*100.

In [12]:
```
#Converting Type of Values present in Column "GamesWon" to Integer.
laliga['GamesWon'] = laliga['GamesWon'].astype(int)

#Converting Type of Values present in Column "GamesPlayed" to Integer.
laliga['GamesPlayed'] = laliga['GamesPlayed'].astype(int)

#Create a new column with the name “Winning_Percent”.
#Calculate Winning Percentage for each Team and append the new Column to original DataFrame.
laliga['Winning_Percent'] = (laliga['GamesWon']/laliga['GamesPlayed']) *100

#If there are any numerical error or Null Values then replace it with 0%
laliga['Winning_Percent'].fillna(0,inplace = True)

#Print the Team Name and its Winning Percentage.
laliga[['Team','Winning_Percent']]
```

Out[12]:

	Team	Winning_Percent
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

Task 7: **Print the top 5 teams which have the highest Winning percentage (5 points)**

In [13]:
```
#Creating a new dataframe containing teams with their Winning Percentage
df_win_per = laliga[['Team','Winning_Percent']].copy()

#Converting Type of Values present in Column "Winning_Percentage" to Float.
df_win_per['Winning_Percent'] = df_win_per['Winning_Percent'].astype(float)

#Sorting the Dataframe "df_win_per" on values of Winning Percentage
df_win_per.sort_values(by='Winning_Percent', ascending=False, inplace=True)

#Printing the top 5 teams which have the highest Winning Percentage
df_win_per.head(5)
```

Out[13]:

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

Task 8: **Group teams based on their “Best position” and print the sum of their points for all positions (15 points)**

Eg: Best Position Points

```
1    25000

2    7000
```

In [14]:
```
#Converting Type of Values present in Column "Points" to Integer.
laliga['Points'] = laliga['Points'].astype(int)

#Converting Type of Values present in Column "BestPosition" to Integer.
laliga['BestPosition'] = laliga['BestPosition'].astype(int)

#Group teams based on their “Best position”
```

```
df_group_BestPosition = laliga[['Team','Points','BestPosition']].groupby('BestPosition')

#Compute sum of group values on BestPosition and Print them
df_group_BestPosition.sum()
```

Out[14]:

BestPosition	Points
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

Mini Project 3 Finished !!!

Submitted By: Raunak Choudhary

Contact: raunakchoudhary17@gmail.com