Project - Python

Name: Panagam Mohitha

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

Attribute Information:

- **Pos** - Position in among the list of all teams
- **Team Seasons** - how many seasons team has played so far
- **Points** - total number of points of the team
- **GamesPlayed**- total number of games played so far
- **GamesWon**- total number of games won so far
- **GamesDrawn** - total number of games drawn so far
- **GamesLost** - total number of games lost so far
- **GoalsFor** - total number of goals by the team
- **GoalsAgains** - total number of goals against the team
- **Champion** - total number of times it team is a champion
- **Runner-up** - total number of times it team is a runner-up
- **Third / Fourth/ Fifth/ Sixth** - total number of times it team came in a third/fourth/fifth/sixth position
- **Debut** - debut year
- **BestPosition** - best position of the team

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.

Task-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)

Importing the library

In [101…]
```
import numpy as np
import pandas as pd
from matplotlib import pyplot as plt
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.api as sm
import scipy.stats as stats
from scipy.stats import chisquare,chi2_contingency
from scipy.stats import ttest_ind
from scipy.stats import f_oneway
from sklearn.preprocessing import LabelEncoder
import warnings
warnings.filterwarnings("ignore")
import copy
```

Read data as Data frame:

In [102…]
```
laliga_scores = pd.read_csv('C:\\Users\\Mohitha Panagam\\Downloads\\PROJECT\\Laliga
```

In [103…]
```
laliga_scores.head(10)
```

Out[103]:

	Pos	Team	Seasons	Points	GamesPlayed	GamesWon	GamesDrawn	GamesLost	GoalsFo
0	1	Real Madrid	86	4385	2762	1647	552	563	594
1	2	Barcelona	86	4262	2762	1581	573	608	590
2	3	Atletico Madrid	80	3442	2614	1241	598	775	453
3	4	Valencia	82	3386	2664	1187	616	861	439
4	5	Athletic Bilbao	86	3368	2762	1209	633	920	463
5	6	Sevilla	73	2819	2408	990	531	887	368
6	7	Espanyol	82	2792	2626	948	608	1070	360
7	8	Real Sociedad	70	2573	2302	864	577	861	322
8	9	Zaragoza	58	2109	1986	698	522	766	268
9	10	Real Betis	51	1884	1728	606	440	682	215

Shape of the data

In [104…]
```
laliga_scores.shape
```

Out[104]:
```
(61, 18)
```

Inference:

- There are 61 Observations / Rows and 18 Attributes / Columns.

Data type of each attribute

In [105…]
```
laliga_scores.info()
laliga_scores.isnull().sum()
```

```
<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
```

Out[105]:
```
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
Debut           0
BestPosition    0
dtype: int64
```

Inference:

- Here The 'attribute Points' 'GamesPlayed' 'GamesWon' 'GamesDrawn' 'GamesLost' 'GoalsFor' 'Goals' 'Champion' 'Runner-up' 'Third' 'Fourth' 'Fifth' 'Sixth' and 'Debut' are of

type object i.e categorical variables.

- Rest all other attributes are of int type.
- We could also see there are no missing values found in the data.
- Also, all the attributes have no-null data.

Replace '-' with '0'

In [106…
```
laliga_scores.replace('-', 0, inplace=True)
laliga_scores
```

Out[106]:

	Pos	Team	Seasons	Points	GamesPlayed	GamesWon	GamesDrawn	GamesLost	GoalsF
0	1	Real Madrid	86	4385	2762	1647	552	563	59
1	2	Barcelona	86	4262	2762	1581	573	608	59
2	3	Atletico Madrid	80	3442	2614	1241	598	775	45
3	4	Valencia	82	3386	2664	1187	616	861	43
4	5	Athletic Bilbao	86	3368	2762	1209	633	920	46
...	...	...	...	...	...	...	...	...	
56	57	Xerez	1	34	38	8	10	20	
57	58	Condal	1	22	30	7	8	15	
58	59	Atletico Tetuan	1	19	30	7	5	18	
59	60	Cultural Leonesa	1	14	30	5	4	21	
60	61	Girona	1	0	0	0	0	0	

61 rows × 18 columns

Inference:

- Replace function is used to replace '-' with '0' which will allow us for arithematic operations

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

In [107…
```
laliga_scores['Debut'] = laliga_scores['Debut'].astype(str)
```

```
Debut_Year = laliga_scores[laliga_scores['Debut'].str[:4].between('1930','1980')]

Debut_Year[['Team','Debut']]
```

Out[107]:

	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

	Team	Debut
59	Cultural Leonesa	1955-56

Inference:

- 37 teams debuted between 1930-1980, including 1930 excluding 1980.
- Valencia in early 30's and cultural leonesa in latest 50's.

Task-3 Print the list of teams which came Top 5 in terms of points

In [108…
```
laliga_scores_sort = laliga_scores[['Team','Points']].copy()

laliga_scores_sort['Points'] = laliga_scores_sort['Points'].astype(int)

laliga_scores_sort.sort_values(by='Points', ascending=False, inplace=True)

laliga_scores_sort.head(5)
```

Out[108]:

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

Inference:

- Sorted top teams using sort function and extracted top 5 using head function
- Real Madrid with 4385 top all the teams
- Followed by Barcelona, Atletico Madrid, Valencia, Athletic Bilbao teams

Task-4 Write a function with the name “Goal_diff_count” which should return all the teams with their Goal Differences.

- **Goal_diff_count = GoalsFor - GoalsAgainst**

In [109…
```
laliga_scores['GoalsFor'] = laliga_scores['GoalsFor'].astype(int)
laliga_scores['GoalsAgainst'] = laliga_scores['GoalsAgainst'].astype(int)

def Goal_diff_count():
     laliga_scores['Goal_diff_count'] = laliga_scores['GoalsFor']-laliga_scores['G
     return laliga_scores[['Team','Goal_diff_count']]

Goal_score = Goal_diff_count()

Goal.sort_values(by = 'Goal_diff_count',ascending=False)
```

Out[109]:

	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

Inference:

- Created a function 'Goal_diff_count()' which return all the teams with Difference between Goal for and Goal aganist.
- Real Madrid has highest goal difference count
- Racing Santander has the least count.

Task-5 Using the same function, find the team which has a maximum and minimum goal difference.

In [110…
```
Goal_diff_count().head(1)
```

Out[110]:

	Team	Goal_diff_count
0	Real Madrid	2807

In [111…
```
Goal_diff_count().tail(1)
```

Out[111]:

	Team	Goal_diff_count
60	Girona	0

Inference:

- Real Madrid has highest goal difference count with 2807 difference.
- Girona has the least with 0 goal difference.

Task-6 Create a new column with the name “Winning Percent” and append it to the data set

- **Percentage of Winning = (GamesWon / GamesPlayed)*100**
- **If there are any numerical error, replace it with 0%**

In [112…
```
laliga_scores['GamesWon'] = laliga_scores['GamesWon'].astype(int)
laliga_scores['GamesPlayed'] = laliga_scores['GamesPlayed'].astype(int)

laliga_scores['Winning Percent'] = (laliga_scores['GamesWon']/laliga_scores['GamesI

laliga_scores['Winning Percent'].fillna(0,inplace = True)

laliga_scores[['Team','Winning Percent']]
```

Out[112]:

	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

In [113…
```
laliga_scores.shape
```

Out[113]:
```
(61, 20)
```

Inference:

- Real Madrid has highest winning percent of 59.63%.
- Girona has the least with 0 %
- Two new columns added to the data frame so that the shape is changed from (61, 18) to (61, 20)

Task-7 Print the top 5 teams which have the highest Winning percentage

In [114…
```
laliga_scores[['Team','Winning Percent']].head(5)
```

Out[114]:

	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

Inference:

- Real Madrid has highest winning percent of 59.63%.
- Followed by Barcelona, Atletico Madrid, Valencia, Athletic Bilbao teams

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

Eg: Best Position Points

1 25000

2 7000

In [115…

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

# grouping teams based on 'BeastPosition' column
Best_group = laliga_scores[['Team','Points','BestPosition']].groupby('BestPosition

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

Out[115]:

```
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
```

Inference:

- The best of group points are of 27933 is the highest whereas 34 is least.