Propuesta de Proyecto Final

En este proyecto contamos con el siguiente dataset, que contienen estadisticas de los equipos de la Major League Baseball del año 1871 a 2022.

Librerias necesarias

In [5]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import csv
import matplotlib.pyplot as plt
from sklearn import linear_model
```

Limpieza de Datos

El dataset cuenta con varias columnas que no nos interesan para este proyecto, nos interesan las columnas numericas de las estadisticas por equipo, de igual manera estas columnas cuentan con valores nulos, los cuales llenamos las medianas de los datos.

In [8]:

```
import sqlite3

conn = sqlite3.connect('lahman_1871-2022.sqlite')



query = '''select * from Teams
inner join TeamsFranchises
on Teams.franchID == TeamsFranchises.franchID
where Teams.G >= 150 and TeamsFranchises.active == 'Y';
'''

Teams = conn.execute(query).fetchall()

teams_df = pd.DataFrame(Teams)

cols = ['yearID','lgID','teamID','franchID','divID',
        'Rank','G','Ghome','W','L','DivWin','WCWin',
        'LgWin','WSWin','R','AB','H','2B','3B','HR',
        'BB','SO','SB','CS','HBP','SF','RA','ER','ERA',
        'CG','SHO','SV','IPouts','HA','HRA','BBA','SOA',
        'E','DP','FP','name','park','attendance','BPF',
        'PPF','teamIDBR','teamIDlahman45','teamIDretro',
        'franchID','franchName','active','NAassoc']


teams_df.columns = cols



drop_cols = ['lgID','franchID','divID','Rank','Ghome',
             'L','DivWin','WCWin','LgWin','WSWin','SF',
             'name','park','attendance','BPF','PPF',
             'teamIDBR','teamIDlahman45','teamIDretro',
             'franchID','franchName','active','NAassoc']
df = teams_df.drop(drop_cols,axis=1)

print("Datos nulos por columna: ")
print(df.isnull().sum(axis=0).tolist())


df = df.drop(['CS','HBP'],axis=1)

df['SO'] = df['SO'].fillna(df['SO'].median())
df['DP'] = df['DP'].fillna(df['DP'].median())

print("Datos nulos después de los cambios:")
print(df.isnull().sum(axis=0).tolist())
```

```
Datos nulos por columna:
[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 16, 0, 418, 943, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0]
Datos nulos después de los cambios:
[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0]
```

In [11]:
```
1 df
```

Out[11]:

	yearID	teamID	G	W	R	AB	H	2B	3B	HR	BB	SO	SB	RA	ER	ER
0	1961	LAA	162	70	744	5424	1331	218	22	189	681	1068.0	37	784	689	4.3
1	1962	LAA	162	86	718	5499	1377	232	35	137	602	917.0	46	706	603	3.7
2	1963	LAA	161	70	597	5506	1378	208	38	95	448	916.0	43	660	569	3.5
3	1964	LAA	162	82	544	5362	1297	186	27	102	472	920.0	49	551	469	2.9
4	1965	CAL	162	75	527	5354	1279	200	36	92	443	973.0	107	569	508	3.1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	.
2431	2017	WAS	162	97	819	5553	1477	311	31	215	542	1327.0	108	672	623	3.8
2432	2018	WAS	162	82	771	5517	1402	284	25	191	631	1289.0	119	682	649	4.0
2433	2019	WAS	162	93	873	5512	1460	298	27	231	584	1308.0	116	724	683	4.2
2434	2021	WAS	162	65	724	5385	1388	272	20	182	573	1303.0	56	820	743	4.8
2435	2022	WAS	162	55	603	5434	1351	252	20	136	442	1221.0	75	855	785	5.0

2436 rows × 27 columns

Descripción de Estadísticas de nuestro dataset

- **G:** Número de juegos jugados por el equipo.
- **W:** Número de victorias del equipo.
- **R:** Número total de carreras anotadas por el equipo.
- **AB:** Número total de turnos al bate (At Bats) realizados por el equipo.
- **H:** Número total de hits (batazos seguros) conectados por el equipo.
- **2B:** Número total de dobles conectados por el equipo.
- **3B:** Número total de triples conectados por el equipo.
- **HR:** Número total de cuadrangulares (home runs) conectados por el equipo.
- **BB:** Número total de bases por bolas otorgadas por el equipo.
- **SO:** Número total de ponches (strikeouts) realizados por los bateadores del equipo.
- **SB:** Número total de bases robadas por el equipo.
- **RA:** Número total de carreras permitidas por la defensa del equipo.
- **ER:** Número total de carreras limpias permitidas por la defensa del equipo.
- **ERA:** Promedio de carreras limpias permitidas por cada 9 entradas lanzadas por el equipo.
- **CG:** Número total de juegos completos lanzados por los lanzadores del equipo.
- **SHO:** Número total de blanqueadas (shutouts) realizadas por el equipo.
- **SV:** Número total de partidos salvados por los relevistas del equipo.
- **IPouts:** Número total de outs registrados por los lanzadores del equipo.
- **HA:** Número total de hits permitidos por la defensa del equipo.
- **HRA:** Número total de cuadrangulares permitidos por la defensa del equipo.
- **BBA:** Número total de bases por bolas otorgadas por los lanzadores del equipo.

- **SOA:** Número total de ponches (strikeouts) realizados por los lanzadores del equipo.
- **E:** Número total de errores cometidos por la defensa del equipo.
- **DP:** Número total de doble plays realizados por la defensa del equipo.
- **FP:** Porcentaje de fildeo (Fielding Percentage) del equipo.

Propuestas de Proyecto

Nos interesa calcular estadisticas avanzadas en el Béisbol como:

1. Slugging Percentage (SLG):

La SLG mide la capacidad de un bateador para producir bases adicionales con sus batazos.

$$SLG = \frac{1B + 2 \times 2B + 3 \times 3B + 4 \times HR}{AB}$$

- (1B:) Número de sencillos (hits de una base).
- (2B:) Número de dobles.
- (3B:) Número de triples.
- (HR:) Número de cuadrangulares.
- (AB:) Número de turnos al bate.

2. On-Base Percentage (OBP):

El OBP mide la capacidad de un bateador para llegar a base y evitar ser eliminado.

$$OBP = \frac{H + BB + HBP}{AB + BB + HBP + SF}$$

- (H:) Número total de hits.
- (BB:) Número total de bases por bolas otorgadas.
- (HBP:) Número total de veces que el bateador es golpeado por un lanzamiento.
- (AB:) Número total de turnos al bate.
- (SF:) Número total de sacrificios de fly.

Intereses:

1. Hacer exploración de los datos usando visualizaciones informativas que permiten entender mejor cada una de las estadisticas del juego

2. Analizar la correlación que existe entre cada una de las estadisticas tanto clasicas como avanzadas y la cantidad de victorias que logra un equipo en una temporada.

3. Crear un modelo de predicción (regresión lineal multivariada) para predecir la cantidad de victorias por equipo.

Como el dataset contiene datos hasta la temporada 2022, podemos predecir las victorias para la temporada 2023 y como esta temporada ya termino podemos evaluar con los resultados reales que tan bien funciona nuestro modelo. Además de predecir la temporada 2024 que esta