UNIVERSIDAD CENTRAL DEL ECUADOR

PROGRAMACIÓN PARA DISPOSITIVOS MÓVILES

Título:

Ejercicio en Clase

Nombre: Elvis Herrera

Semestre: Décimo

Fecha: 19-12-2023

Programación para Dispositivos Móviles - Web Scraping

Web Scraping Mediante API:

Web scraping mediante API implica la utilización de las APIs proporcionadas por sitios web para obtener datos de manera más directa y estructurada. En lugar de analizar el HTML de una página web, se realizan solicitudes a la API del sitio, que responde con datos en un formato específico. Este enfoque es preferible cuando un sitio web ofrece una API, ya que proporciona una manera más eficiente, estructurada y generalmente más amigable con los desarrolladores para acceder a la información.

Ejercicio:

Para el presente ejercicio, realizamos web scraping en la página de wikiwand.com, centrándonos en buscar información sobre la Premier League. En particular, seleccionamos la tabla de 'Clasificación histórica de la Premier League' para extraer la información relevante.

Enlace de la página:

https://www.wikiwand.com/es/Premier_League#Clasificaci%C3%B3n_hist%C3%B3rica_Premier_League

Enlace de la API de la tabla:

https://www.wikiwand.com/mcs-api/es.wikipedia.org/v1/page/mobile-sections-remaining/Premier_League

Enlace documentos Git Hub:

https://github.com/emherrerat/Deber_Dispositivos

En el desarrollo de este proyecto, se llevaron a cabo dos procesos principales utilizando archivos con extensión ipynb y una base de datos obtenida mediante web scraping del sitio mencionado. Los archivos ipynb contienen el código y la lógica utilizada durante la tarea, y fueron compartidos y almacenados en el repositorio de GitHub para facilitar el acceso y la colaboración.

La tarea involucró la extracción de datos específicos relacionados con la Premier League mediante web scraping del sitio web mencionado. La información recopilada se organizó y guardó en un archivo CSV denominado 'PREMIER.csv'. Este archivo CSV sirve como una fuente estructurada de datos, facilitando su análisis y manipulación.

Análisis de la Data

Estadísticos básicos para comprender la data

	PJ	PG	PE	PP	G. Favor	G. contra	Dif	Puntos	Títulos	Temporadas
count	10.000000	10.000000	10.00000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000
mean	3921.800000	1662.900000	960.20000	1298.700000	6277.200000	5383.000000	894.200000	5948.900000	9.500000	98.800000
std	428.067959	269.231602	119.82005	163.640561	824.634303	552.068836	707.596213	914.781874	6.059886	11.409548
min	3340.000000	1260.000000	780.00000	1075.000000	5143.000000	4926.000000	22.000000	4560.000000	2.000000	86.000000
25%	3566.500000	1463.750000	886.00000	1195.500000	5494.250000	5018.250000	518.750000	5289.250000	6.000000	88.750000
50%	3851.000000	1630.500000	940.50000	1266.500000	6417.500000	5187.000000	699.000000	5835.500000	8.000000	96.000000
75%	4270.000000	1887.500000	1044.75000	1334.500000	7045.500000	5361.750000	1499.250000	6779.500000	12.000000	107.500000
max	4670.000000	2035.000000	1171.00000	1613.000000	7195.000000	6482.000000	2040.000000	7166.000000	20.000000	120.000000

Interpretación de la matriz

- **Count (Recuento):** Representa el número de clubes en cada columna, en la matriz podemos observar que se tienen datos para los 10 clubes en todas las columnas.
- **Mean (Media):** Es el promedio aritmético de los valores en cada columna, en la matriz el promedio de partidos jugados (PJ) es aproximadamente 3921.8.
- **Std (Desviación estándar):** Mide la dispersión de los valores con respecto a la media es decir cuánto varían los valores de la media, en la matriz la desviación estándar de partidos jugados (PJ) es aproximadamente 428.07.
- **Min (Mínimo):** Representa el valor mínimo en cada columna, en la matriz el club con menos partidos jugados (PJ) tiene 3340.
- **25%, 50%, 75% (Denominados Percentiles):** Son los valores que dividen el conjunto de datos en percentiles, en la matriz observamos que el 25% de los clubes tienen menos de 3566.5 partidos jugados.
- **Max (Máximo):** Representa el valor máximo en cada columna, en la matriz se observa que el club con más partidos jugados (PJ) tiene 4670.

Matriz de Correlación

	PJ	PG	PE	PP	G. Favor	G. contra	Dif	Puntos	Títulos	Temporadas
PJ	1.00	0.87	0.97	0.47	0.94	0.73	0.54	0.89	0.50	0.99
PG	0.87	1.00	0.88	-0.01	0.97	0.32	0.88	1.00	0.80	0.81
PE	0.97	0.88	1.00	0.36	0.91	0.62	0.58	0.91	0.51	0.94
PP	0.47	-0.01	0.36	1.00	0.21	0.93	-0.47	0.03	-0.39	0.57
G. Favor	0.94	0.97	0.91	0.21	1.00	0.53	0.75	0.97	0.70	0.91
G. contra	0.73	0.32	0.62	0.93	0.53	1.00	-0.16	0.36	-0.08	0.80
Dif	0.54	0.88	0.58	-0.47	0.75	-0.16	1.00	0.85	0.89	0.44
Puntos	0.89	1.00	0.91	0.03	0.97	0.36	0.85	1.00	0.77	0.84
Títulos	0.50	0.80	0.51	-0.39	0.70	-0.08	0.89	0.77	1.00	0.45
Temporadas	0.99	0.81	0.94	0.57	0.91	0.80	0.44	0.84	0.45	1.00

Interpretación de la matriz

En la matriz de correlación nos muestra los siguientes resultados:

- **Correlación entre Partidos Jugados (PJ) y las otras variables:**

Tenemos una alta correlación positiva con Temporadas (0.99) y Puntos (0.89), una correlación positiva fuerte con Goles a Favor (0.94) y Títulos (0.50).

- **Correlación entre Partidos Ganados (PG) y las otras variables:**

Tenemos una alta correlación positiva con Puntos (1.00) y Goles a Favor (0.97), una correlación positiva fuerte con Diferencia de Goles (0.88).

- **Correlación entre Partidos Empatados (PE) y las otras variables:**

Tenemos una alta correlación positiva con Partidos Jugados (0.97) y Goles a Favor (0.91), una correlación positiva fuerte con Títulos (0.51).

- **Correlación entre Partidos Perdidos (PP) y otras variables:**

Tenemos una correlación negativa moderada con Diferencia de Goles (-0.47) y una correlación positiva moderada con Goles en Contra (0.93).

- **Correlación entre Goles a Favor y otras variables:**

Tenemos una alta correlación positiva con Goles en Contra (0.53), Puntos (0.97), y Partidos Jugados (0.94).

- **Correlación entre Goles en Contra y otras variables:**

Tenemos una alta correlación positiva con Partidos Perdidos (0.93) y moderada con Diferencia de Goles (0.36).

- **Correlación entre Diferencia de Goles y otras variables:**

Tenemos una correlación positiva fuerte con Goles a Favor (0.75) y Partidos Jugados (0.54), una correlación negativa moderada con Partidos Perdidos (-0.47).

- **Correlación entre Puntos y otras variables:**

Tenemos una alta correlación positiva con Partidos Jugados (0.89) y Partidos Ganados (1.00) y una correlación positiva fuerte con Goles a Favor (0.97) y Diferencia de Goles (0.85).

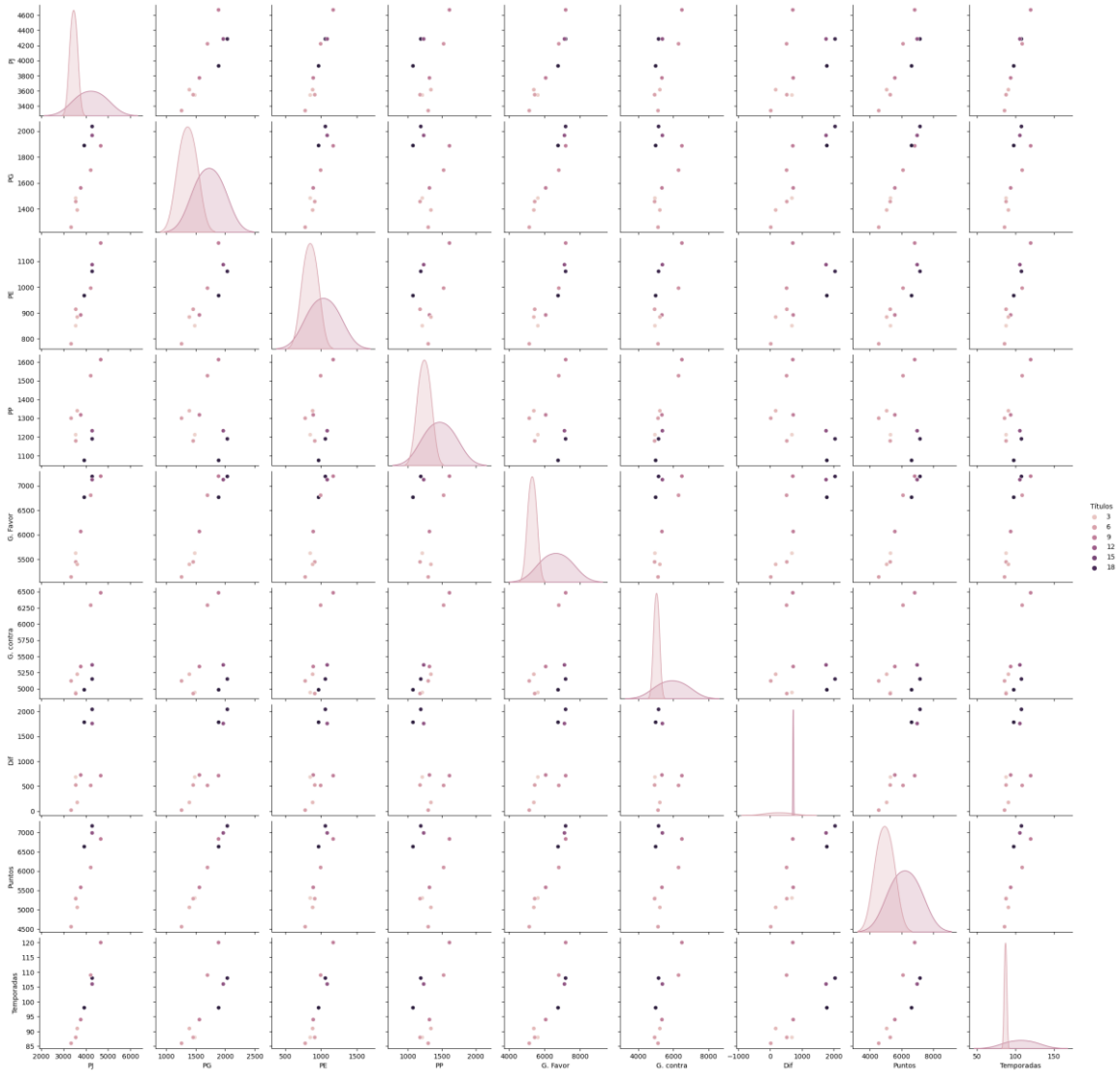
- **Correlación entre Títulos y otras variables:**

Tenemos una correlación positiva moderada con Partidos Ganados (0.80) y Diferencia de Goles (0.89).

- **Correlación entre Temporadas y otras variables**:

Tenemos una alta correlación positiva con Partidos Jugados (0.99) y Puntos (0.84) y una correlación positiva fuerte con Goles a Favor (0.91) y Partidos Empatados (0.94).

Grafica de correlaciones respecto a los Estados



Interpretación de la matriz

En el grafico podemos observar cómo los datos están medianamente dispersos, es decir que las observaciones en el conjunto de datos están ampliamente distribuidas en torno a la media, sin seguir un patrón claro o predecible. La dispersión nos indica la variabilidad que existe en este conjunto de datos. Lo que indica que las observaciones son más homogéneas.

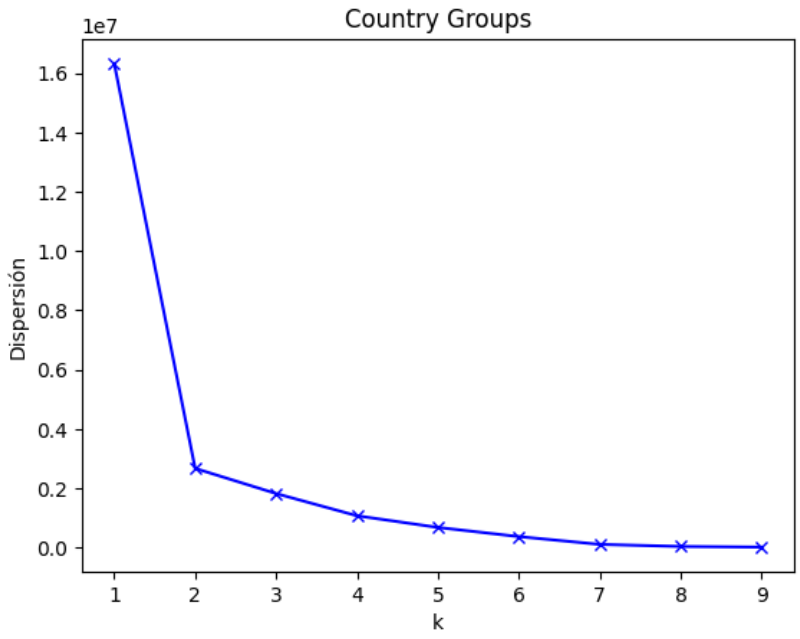
Selecciónanos de variables significativas

```
sel_data = data[['PJ','PG', 'PE', 'PP', 'G. Favor', 'Puntos', 'Títulos']]
```

Y obtenemos la siguiente tabla:

	PJ	PG	PE	PP	G. Favor	Puntos	Títulos
0	4286	2035	1061	1190	7190	7166	19
1	4286	1966	1087	1233	7125	6985	13
2	4670	1886	1171	1613	7195	6829	9
3	3930	1888	967	1075	6766	6631	20
4	4222	1699	996	1527	6807	6093	7
5	3772	1562	892	1318	6069	5578	9
6	3546	1484	850	1212	5627	5302	2
7	3550	1457	914	1179	5450	5285	6
8	3616	1392	884	1340	5400	5060	4
9	3340	1260	780	1300	5143	4560	6

Análisis de Dispersión por método de Codo



En el gráfico podemos determinar que el número óptimo de clústeres en nuestro conjunto de datos seria **2**.